Defanging facial recognition:

A statistical approach to bias mitigation and policy conditions for responsible use

Stephanie Cairns

Department of Mathematics and Statistics
McGill University, Montreal
June 2021

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

Abstract

Facial recognition (FR) technology threatens privacy and has been shown to exhibit significant bias against certain demographic groups, which can have serious implications when deployed by police or other government agents. This thesis explores conditions that should be met before a widely proposed moratorium on facial recognition technology could be safely lifted. Such conditions include the establishment of an auditing system to measure accuracy and bias in FR models and the implementation of a governance framework for the collection and use of FR data. We also introduce a statistical method that reduces model bias while increasing accuracy, all without requiring demographic information about model subjects. This method can be integrated into current FR models without need for retraining.

Résumé

La technologie de reconnaissance faciale (RF) menace la vie privée et a fait preuve de partialité contre certains groupes démographiques, ce qui peut avoir de graves implications lorsque cette technologie est déployée par la police ou par d'autres agents gouvernementaux. Cette thèse explore les conditions qui devraient être remplies avant qu'un moratoire largement proposé sur la technologie de reconnaissance faciale puisse être levé en toute sécurité. Ces conditions comprennent la mise en place d'un système de vérification pour mesurer la précision et le biais des modèles RF, ainsi que la mise en œuvre d'un cadre de gouvernance pour la collecte et l'utilisation des données RF. Nous introduisons également une méthode statistique qui réduit le biais d'un modèle tout en augmentant sa précision, sans nécessiter d'informations démographiques sur les sujets du modèle. Cette méthode peut être intégrée à un modèle RF actuel sans ré-entraîner le modèle.

Acknowledgements

Firstly, I would like to thank my supervisor Adam Oberman for his support and flexibility in allowing me to undertake a rather unorthodox master's thesis, one that combines math and machine learning with public policy. Thank you, Adam, for all the help you have given me over the course of this project and over the course of my degree—I feel very privileged to have been your student.

A core component (Chapter 3) of this thesis was inspired by a policy project [66, 67] I completed in collaboration with Sonja Solomun, Sara Parker, Ellen Rowe, and Charlotte Reboul, under the supervision of Derek Ruths and Taylor Owen[1]. Thanks for all your hard work and enthusiastic support, and thanks for showing me that machine learning and public policy are not so unrelated as they may at first appear.

Next, a huge thank you goes to Tiago Salvador for embracing my interest in facial recognition bias and for taking the lead on a paper [83] that would become a key part of this thesis. Chapter 4 details (in my own words) the bias mitigation method presented in our paper. As first author, Tiago played a large role in designing the method. He also coded and ran our experiments and created the tables and figures that appear in Chapter 4. Tiago, working with you has been a pleasure, and I greatly appreciate all the help you've given me. Thank you also to Vikram Voleti, who provided guidance for the project, and to Noah Marshall, who helped write part of the code needed to run the experiments.

[1]Note that the work in Chapter 3 is my own and that the aforementioned team played no direct role in producing this thesis.

Finally, where would I be without my loving, aggravating, and endlessly supportive family? In the not-so-wise words of my father, “a master’s degree is the new high school diploma" (no pressure, right?), so thank you for helping me survive an only slightly less gruelling second high school experience. I would also like to thank my dog Zeus for setting aside his obsession with toy foxes to provide the emotional support I needed to finish this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Facial recognition (FR) technology has become pervasive and virtually inescapable: it has infiltrated Canadian airports [60], malls [97], and, prior to a recent investigation by the Office of the Privacy Commissioner, both municipal police departments and the Royal Canadian Mounted Police (RCMP), Canada's federal police service [63]. Despite an explosion in the use of FR, particularly by government agencies, Canada's data, privacy, and technology laws have remained stagnant and have largely failed to adequately regulate the powerful new technology [94]. FR systems have many potential benefits and have been successfully employed by police departments to find missing children [50] and to identify child victims of sexual abuse [35]. These successes, however, are undercut by the significant threat that FR, especially when deployed in public places, poses to privacy and meaningful consent [7]. The technology has also been shown to exhibit bias against certain demographic groups, particularly Black people [11, 19, 29, 43, 71]; this technological bias, amplified by the human bias already present in law enforcement [70], risks causing serious harm to marginalized groups. Already, FR systems have led to the false arrests of multiple Black men in the United States [33]. Many major American cities have responded to these concerns by imposing moratoriums or bans on facial recognition technology [77]; in Canada, there have been growing calls for the federal government to do the same [94].

This thesis examines two central questions: firstly, given a hypothetical federal moratorium on the use of FR by Canadian government agencies, which conditions, both technical and policy-related, would need to be put in place before such a moratorium could safely and ethically be lifted? And secondly, how can we start working toward fulfilling one of these conditions, namely the condition of context-based bias mitigation?

To tackle the first question, we consider fundamental research problems that must be solved, as well as changes to Canadian laws that must be made and policies that must be implemented before the risks posed by FR may be considered sufficiently dampened. Public policy alone will not solve the problem of facial recognition; neither, however, will research and technical solutions. We cannot program our way to a free and equitable society; for instance, even if highly effective bias mitigation solutions become widely known, regulations are required to ensure that the FR systems used by police officers, border control agents, and other government employees are indeed unbiased. Conversely, sound policies must be informed by research: successfully measuring bias necessitates an understanding of the nuances and mutual incompatibility of different bias metrics. Employing a simple, one-size-fits-all metric entirely ignores the diverging contexts in which the tool may be deployed: for instance, equal false positive rates should be prioritized in situations where false positive errors (e.g. false arrests) may pose significant risks, while in other cases equal false negative errors may be more important. Technical, legal, and policy experts must therefore work in concert to address the problem of bias. The same is true for the two other groups of conditions—related to data and FR usage—that we explore.

1.1 Summary

We begin Chapter 2 with a brief explanation of facial recognition (FR) technology, its uses in Canada, and its most severe shortcomings. We then include a survey of the various mathematical definitions of fairness, followed by a demonstration of the mutual incompatibility of three of these definitions.

Chapter 3 introduces the implications of a federal moratorium on government uses of FR (section 3.1), then dives into the conditions necessary for removing such a moratorium. Note that these conditions are non-exhaustive and fall into three primary camps: accuracy and bias conditions, data conditions, and usage conditions, with the greatest attention paid to the first category. In section 3.2, we consider how FR researchers have historically defined bias, how bias and accuracy should in fact be measured (i.e. by accounting for the context and the use case), which testing datasets should be used when measuring bias and accuracy, which demographic groups should be prioritized, and what factors cause bias. We conclude the section with a thorough investigation of bias mitigation strategies. Sections 3.3 addresses data conditions and provides an overview of current data and privacy laws before introducing key considerations for developing more appropriate data regulations. Similarly, section 3.4 presents considerations for establishing conditions and limits on the use of FR. These conditions include assessing the risk, particularly at the societal level, of various FR use cases, as well as assessing use cases' proportionality (i.e. whether employing an FR solution is disproportionate to a given problem).

Finally, Chapter 4 delves into a bias mitigation approach developed by Salvador et al. [83], to which I contributed. Our method, which applies the under-utilized fairness definition of fairness-calibration to FR, could be used with any FR system with no re-training or demographic data labels required, the latter of which are often non-existent. Most importantly, it would allow FR users to select, based on the deployment context, which definitions of fairness they wish to prioritize, thereby helping to satisfy the condition of context-dependent bias mitigation.

Chapter 2

Background

2.1 Facial recognition technology

2.1.1 How it works

Current state-of-the-art facial recognition models are composed of deep neural networks that accept images of face images as input and that output face embeddings—vectorized approximations of facial features. In order to output accurate face embeddings, a model is first trained on vast quantities of facial images, which contain identity labels (e.g. 'John Smith'). Training involves the use of a loss function, which at each training interval evaluates how well the model is performing on a set of training images.

Facial recognition technology can be used to achieve two different goals: one-to-many identification and one-to-one verification. The former involves scanning a large database for images that match a target image, while the latter involves checking whether the target image truly matches another photo; this second type of FR may be used to unlock one's cell phone, for example. Both forms of FR compare the face embeddings of images to determine whether they depict the same person; more precisely, a pair of images x_i and x_j is deemed a match if the cosine similarity score

of their respective embeddings $f(x_i)$ and $f(x_j)$ exceeds a particular threshold. Their cosine similarity score is computed as $s(x_i, x_j) = \frac{f(x_1) \cdot f(x_2)}{\|f(x_1)\| \|f(x_2)\|}$. A pair of embeddings with a high cosine similarity score can be assumed to be highly similar to one another.

2.1.2 Recent and current uses

Major developers

While many prominent technology companies like Facebook develop facial recognition algorithms solely for their own use (e.g. in automatically tagging user photos [84]), numerous other companies sell their FR technology to law enforcement, private businesses, and other buyers. Notable developers include Amazon, which made headlines in June 2020 when it declared a one-year moratorium on sales of their FR algorithm, Rekognition, to law enforcement, though the model is still available for commercial and academic use [103]. Microsoft and IBM took similar measures, with the former imposing an indefinite moratorium until a “national law...grounded in human rights" could be established, and the latter permanently ceasing all production and sales of FR technology [79]. In 2019, the AI ethics board of Axon, a technology and weapons company, likewise concluded that FR technology was not yet reliable enough to be sold to law enforcement [6].

Other companies, however, endeavoured to continue selling FR systems to police departments. Until recently, Clearview AI was the largest provider of FR technology to law enforcement in Canada, allowing police officers to match images to the company’s database of three billion photos, which it had obtained by scraping content from social media platforms, in clear violation of the sites’ terms of service [1]. Clearview AI withdrew from the Canadian market in light of a joint investigation by the Office of the Privacy Commissioner of Canada (OPC) and several provincial privacy commissioners [63].

Unlike the UK, the EU, and California, Canada has no legal “right to be forgotten". Despite this, and despite an initial reluctance on behalf of Clearview AI [21], the company eventually allowed Canadians to request that their photos be removed from their database [20]. The privacy

commissioners' investigation culminated in a February 2021 report (with no legal power) declaring that the company's treatment of personal data violated Canadians' privacy rights [58]. As of May 2021, the OPC has not yet concluded a separate but related investigation into the RCMP's use of Clearview AI's FR technology [64, 65].

Other companies continued to operate in Canada; NEC's NeoFace Reveal has been used by Canadian police departments, including the Calgary Police Department, which has used it since 2014 [55]. Idemia, whose algorithm has been shown to exhibit strong racial bias, offered their services to the Sûreté du Québec in 2020 [73].

Canadian Border Protection

Travellers arriving at Canada's major airports are offered the convenient option of checking in at self-serve customs kiosks set up by the Canadian Border Services Agency [60]. These kiosks verify the travellers' identities using facial recognition. Passport Canada also frequently uses FR technology to detect fraud in passport applications [60].

Police

At least 34 police departments across Canada have used Clearview AI's FR tool [1], including Calgary [26], Halifax [75], and Toronto [52] police departments, all three of which initially denied using the technology. The RCMP's National Child Exploitation Crime Centre used Clearview AI's FR technology; other RCMP units briefly employed the technology on a trial basis [81]. The RCMP likewise denied employing Clearview AI's tool before admitting, weeks later, that it had used it for months [98].

Private sector

Private corporations, including retailers like Canadian Tire [104] and Saks Fifth Avenue [27], have also adopted FR technology. One recent high-profile case involved Cadillac Fairview [97], the owner of several of Canada's largest malls. An investigation by the B.C. and Alberta privacy

commissioners found that the real estate company had installed hidden cameras and collected over five million images of shoppers, which were then used to conduct facial recognition.

Fusion systems

While this thesis' focus is on FR, it is important to note that many identification systems, particularly those already deployed in non-Western countries, rely on multiple types of biometric data, such as fingerprints, irises, and gait, in addition to facial images [23]. It is imperative that all Canadian regulations targeting FR also apply to these fusion systems.

2.1.3 Bias

Facial recognition models have been shown to exhibit bias against racial minorities, women, and certain age groups such as children and the elderly. Cook et al. [19] tested 11 commercial FR systems and found that lower skin reluctance (a proxy for a darker skin tone) was associated with lower accuracy. Krishnapriya et al. [43] studied four FR algorithms, finding higher false positive rates and lower false negative rates among Black subjects than white subjects.

In 2019, the US National Institute of Standards and Technology (NIST) [29] undertook a large-scale study of 189 FR algorithms from 99 developers. They tested the algorithms on 18.27 million photos from four datasets: domestic mugshots, photos of immigrants applying for benefits, visa application photos, and border crossing photos.

They compared false positive rates (FPRs) across demographic groups: when searching the datasets of application photos, they found large discrepancies between countries of origin, with FPRs differing by factors of up to 100. The highest FPRs were observed for West and East African and East Asian subjects, with the notable exception of Chinese-made algorithms, which exhibited low FPRs on East Asian faces. The algorithms likewise displayed unequal rates when tested on mugshots, with Indigenous people, followed by African Americans and East Asians having the highest FPRs. Across all algorithms and datasets, women were more likely to be misidentified than men, though

the gender gap in FPRs was smaller than the racial gap. Finally, higher FPRs were noted among children and the elderly.

False negative rates (FNRs) also diverged between groups, though these disparities were not consistent across datasets. For example, along with white faces, African American faces had low FNRs on the mugshot dataset compared to Asian and Indigenous faces. However, when testing on the border crossing dataset, high FNRs were noted for individuals hailing from African and Caribbean nations. This inconsistency is likely explained by differences in image quality between the two datasets.

Research on bias in facial recognition systems has for the most part considered only particular fairness metrics, namely false positive and false negative rates. Other definitions of fairness, such as fairness-calibration, which we will introduce in section 2.2, have so far been overlooked.

2.1.4 Additional risks

FR also poses major risks to privacy and undermines an individual's ability to provide meaningful consent. As they stand, Canada's privacy and data laws are ill-equipped to address these dangers [93]. Moreover, certain uses of FR, such as live police surveillance, engender especially acute risks; there are currently no Canadian regulations to assess, curb, or ban such use cases.

2.2 Mathematical definitions of fairness

Researchers have proposed a myriad of notions to quantify fairness, some of which have been shown to be mutually exclusive. In this section, we present a selection of the most noteworthy or popular definitions. These definitions are taken from survey papers by Verma and Rubin [99] and Mehrabi et al. [49]. For a more extensive (but still not exhaustive) list of definitions, see Appendix A.

Each definition involves the following problem setup: given a machine learning model that outputs a score $s(x) \in [0, 1]$ for an input x, we threshold the model at some value α and set

$$d(x) = \begin{cases} 1, & \text{if } s(x) > \alpha \\ 0, & \text{if } s(x) \leq \alpha \end{cases} \tag{2.1}$$

If $d(x) = 1$, we say that x has been assigned to the positive class. In the case of facial recognition, this corresponds to inputting a pair of face images and outputting a score s and a corresponding decision $d = \{0, 1\}$, predicting whether the pair is truly a match. An alternative example would be a model that predicts whether an individual will be granted a bank loan.

We test a model's level of bias with respect to a particular sensitive attribute G (e.g. race, gender, age group) by dividing subjects into groups based on that attribute. For simplicity, the binary versions of these definitions—where subjects fall into one of two demographic groups $G = \{a, b\}$—are presented. However, the definitions can also be applied to cases with multiple groups.

We can differentiate between two forms of discrimination:

Definition 2.2.1 (Disparate treatment)**.** Discrimination that results from the explicit use of a sensitive attribute (e.g. a person's race, gender, etc.) in making a decision.

Definition 2.2.2 (Disparate impact)**.** Discrimination that does not result from the explicit use of a sensitive attribute but rather from proxy attributes (e.g. neighbourhood and income as proxies for race).

A model that does not directly rely on information about the subject's gender, race, or other sensitive attributes may nevertheless display significant bias against particular groups.

2.2.1 Basic definitions

Many of the definitions of fairness proposed by researchers (see Appendix A for the more complete list) are related to the notions of true and false positives (and negatives). A subject who has been correctly assigned to the positive class by the model corresponds to a true positive (TP), while one who has been incorrectly assigned corresponds to a false positive (FP). The following simple definitions are key to many fairness definitions.

The true positive rate (TPR):

$$\frac{TP}{TP + FN} \tag{2.2}$$

is the proportion of subjects that belong to the positive class that are correctly assigned to the positive class.

The true negative rate (TNR):

$$\frac{TN}{FP + TN} \tag{2.3}$$

is the proportion of subjects that belong to the negative class that are correctly assigned to the negative class.

The false positive rate (FPR):

$$\frac{FP}{FP + TN} = 1 - TNR \tag{2.4}$$

is the proportion of subjects that belong to the negative class that are incorrectly assigned to the positive class.

The false negative rate (FNR):

$$\frac{FN}{TP + FN} = 1 - TPR \tag{2.5}$$

is the proportion of subjects that belong to the positive class that are incorrectly assigned to the negative class.

The positive predictive value (PPV):

$$\frac{TP}{TP + FP} \tag{2.6}$$

is the proportion of positively-assigned subjects that truly belong to the positive class.

2.2.2 Overly simplistic notions of fairness

Definition 2.2.3 (Statistical parity)**.** Also known as demographic parity, group fairness, and equal acceptance rate, statistical parity requires that subjects in both groups a and b have equal probabilities of being assigned to the positive predicted class. That is,

$$P(d = 1|G = a) = P(d = 1|G = b) \tag{2.7}$$

This is equivalent to ensuring that d is independent of G. A severe flaw with this definition is that it can be satisfied in ways that the average person would not deem “fair". For example, for the case of a graduate program with 500 male applicants and 500 female applicants, statistical parity would be satisfied by admitting the top ten percent of female applicants and a random selection of 50 male applicants.

Definition 2.2.4 (Conditional statistical parity)**.** Conditional statistical parity necessitates that subjects in both groups have equal probability of being assigned to the positive predicted class provided they satisfy some legitimate condition L, such as a certain minimum GRE score.[1]

$$P(d = 1|L = l, G = a) = P(d = 1|L = l, G = b) \tag{2.8}$$

[1]Note that this notion of fairness is complicated by the difficulty of identifying a truly “legitimate" condition. GRE scores, for instance, underestimate the success rates of minority and female students [13].

2.2.3 Three fairness definitions of note

Here, $Y = 1$ and $Y = 0$ correspond, respectively, to the events of belonging and not belonging to the positive class.

Definition 2.2.5 (Predictive equality)**.** Predictive equality is satisfied if both groups have equal false positive rates:

$$P(d = 1|Y = 0, G = a) = P(d = 1|Y = 0, G = b) \tag{2.9}$$

Definition 2.2.6 (Equal opportunity)**.** Equal opportunity requires that both groups have equal false negative rates:

$$P(d = 0|Y = 1, G = a) = P(d = 0|Y = 1, G = b) \tag{2.10}$$

Note that this is equivalent to demanding equal true positives rates.

Predictive equality or equal opportunity may be more relevant than the other in particular contexts, depending on whether greater harm might stem from a subject being incorrectly assigned to the positive or to the negative class. Consider an algorithm that decides whether to grant refugee status to asylum seekers (by assessing, for example, the legitimacy of their claim and the risk they face if denied entry to Canada). Guaranteeing that such an algorithm exhibits equal (and low) false negative rates for different groups of applicants is crucial, as a claimant may be subjected to violence and human rights abuses if their application is rejected. Alternatively, if being assigned to the “positive class" by an algorithm carries more risk (e.g. a hypothetical algorithm that determines whether there exists enough evidence to arrest a suspect), equal false positive rates must be prioritized.

In addition to predictive equality and equal opportunity, we introduce a third fairness metric, fairness-calibration, which ensures that, regardless of the group to which a subject (or in the case

of FR, a pair of images) belongs, the score output by the model corresponds to the subject's true probability of being in the positive class. Achieving fairness-calibration allows the users of an algorithm (e.g. police officers, judges, immigration officials) to correctly interpret its scores.

In the case of FR, during testing we compare only pairs of images from the same demographic groups (e.g. we test whether the model can accurately match two white faces, two Black faces, etc.). Here, a "subject" corresponds to a pair of images, which is deemed to belong to the positive class if both images depict the same individual.

Definition 2.2.7 (Fairness-calibration)**.** A model will typically output not only a prediction of the class to which a subject belongs but also a score $S = s$ representing the model's confidence in its prediction. Fairness-calibration, also called well-calibration and fairness within groups, requires that for any given score s, outputted by the model, subjects in both groups with this score have probability s of truly belonging to the positive class.
For any $s \in [0, 1]$:

$$P(Y = 1|S = s, G = a) = P(Y = 1|S = s, G = b) = s \tag{2.11}$$

In other words, the scores outputted by the model should match, regardless of group membership, the true probability of belonging to the positive class.

2.2.4 Incompatible definitions

The case of COMPAS

Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) is a decision support tool developed by Northpointe (now called Equivant) and used by US courts across multiple states, including New York and California. The tool predicts a defendant's probability of re-offending. In 2016, ProPublica [3] published a scathing analysis of COMPAS, highlighting a major discrepancy between the tool's false positive rates for Black and white defendants: 44.9%

and 23.5%, respectively. An inverse discrepancy was found for the tool's false negative rates: 28% for Black defendants and 47.7% for white defendants. In other words, the model was 200% more likely to misclassify non-recidivist Blacks than non-recidivist whites but was 41% less likely to misclassify recidivist Blacks than recidivist whites.

Northpointe published a rebuttal, arguing that ProPublica was incorrect in their assertion that COMPAS was racially biased, as its predictions were fairly-calibrated between groups. The ensuing controversy prompted multiple researchers to explore the relationship between fairness-calibration and other fairness definitions.

Calibration vs Equalized Odds

Chouldechova [12] showed that fairness-calibration and equalized odds[2] (i.e. equal false positive and negative rates between groups) cannot simultaneously hold except under narrow conditions. Kleinberg, Mullainathan, and Raghavan [41] proved a generalized version of that result, namely that the following three conditions: fairness-calibration, balance for the positive class, and balance for the negative class (see Appendix A.2) can only hold if both groups have equal base rates (i.e. the same probability of belonging to the positive class) or in cases of perfect prediction, where the model assigns every subject to the correct class.

We recreate the proof outlined in [12], filling in the gaps and modifying the statement so that it applies to situations with more than two groups. We also stress that the following theorem can be interpreted as a positive and helpful result, allowing ML practitioners to select two of three possible fairness definitions to satisfy: fairness-calibration, equal false positive rates (predictive equality), and equal false negative rates (equal opportunity). The choice of definition will depend on the specific application or problem each practitioner is aiming to tackle. For example, [12] argues for the importance of prioritizing equal false positive and negative rates over fairness-calibration in the case of recidivism algorithms in order to reduce the risk of disparate impact to minority groups.

[2]Note that the notion of equalized odds does not refer to the standard statistical definition of odds (i.e. the probability that an event will occur divided by the probability that the event will not occur).

Theorem 2.2.1. *Let $s(x)$ be the score assigned for each subject x by a machine learning model. Let $G = \{g_1...g_K\}$ be a finite set of mutually exclusive groups such that all $x \in g_i$ for some $i \in \{1...K\}$. Consider the following three fairness definitions: A) fairness-calibration, B) equal false positive rates between groups, and C) equal false negative rates between groups. If A, B, and C are all satisfied, then either perfect prediction holds, or all groups have equal probability of belonging to the positive class, which, due to a wide range of systemic factors affecting education, policing, and other domains, rarely occurs in real life. Otherwise, $s(x)$ may satisfy at most two of the three fairness definitions.*

Proof. Suppose A, B, and C are satisfied. Let d be a function that assigns subjects to the positive class if their outputted score is higher than some arbitrary threshold $\alpha \in [0, 1]$. That is, let

$$d(x) = \begin{cases} 1, & \text{if } s(x) > \alpha \\ 0, & \text{if } s(x) \leq \alpha \end{cases}$$

We first show that each group must have the same positive predictive value (PPV). We then use this fact to prove that satisfying A, B, and C implies that either 1) the probability of truly belonging to the positive class is the same for all groups, or 2) the model achieves perfect prediction ($TPR = 1$ and $FPR = 0$).

Recall that

$$PPV = \frac{TP}{TP + FP} \tag{2.12}$$

which is equivalent to $P(Y = 1|d(x) = 1)$.

Note that since A is satisfied, by definition

$$P(Y = 1|S = s(x), G = g_1) = P(Y = 1|S = s(x), G = g_2) \qquad \forall g_1, g_2 \in G \tag{2.13}$$

Since this holds $\forall s(x) > \alpha$, it follows that

$$P(Y = 1|d(x) = 1, G = g_1) = P(Y = 1|d(x) = 1, G = g_2) \qquad \forall g_1, g_2 \in G \tag{2.14}$$

Thus, the model's PPV is the same for all groups.

Next, let $p_g = P(Y = 1|G = g)$ be the probability that subjects in group g truly belong to the positive class.

We can then see that for any $g \in G$:

$$\begin{aligned} FPR &= \frac{FP}{TN + FP} \\ &= \left(\frac{TP + FN}{TN + FP}\right)\left(\frac{FP}{TP}\right)\left(\frac{TP}{TP + FN}\right) \\ &= \left(\frac{\frac{TP+FN}{TP+FN+TN+FP}}{1 - \frac{TP+FN}{TP+FN+TN+FP}}\right)\left(\frac{1 - \frac{TP}{TP+FP}}{\frac{TP}{TP+FP}}\right)\left(1 - \frac{FN}{TP + FN}\right) \\ &= \left(\frac{p_g}{1 - p_g}\right)\left(\frac{1 - PPV}{PPV}\right)(1 - FNR) \end{aligned} \tag{2.15}$$

Similarly, we can see that for any $g \in G$:

$$\begin{aligned} FNR &= \frac{FN}{TP + FN} \\ &= \left(\frac{TN + FP}{TP + FN}\right)\left(\frac{FN}{TN}\right)\left(\frac{TN}{TN + FP}\right) \\ &= \left(\frac{1 - p_g}{p_g}\right)\left(\frac{1 - NPV}{NPV}\right)(1 - FPR) \end{aligned} \tag{2.16}$$

where $NPV = \frac{TN}{TN+FN}$ is the negative predictive value.

It follows that since all three definitions are satisfied, at least one of the following conditions must hold: either 1) $p_{g_i} = p_{gj}\ \forall g_i, g_j \in G$, meaning that all groups have an equality probability of belonging to the positive class; or 2) $FPR = 0$ and $FNR = 0$, in which case perfect prediction is achieved. □

Chapter 3

Conditions for lifting a facial recognition moratorium

The contents of this chapter are a substantial expansion of two policy articles I co-produced under the supervision of Taylor Owen (Max Bell School of Public Policy) and Derek Ruths (School of Computer Science) [66, 67]. I played a leading role in researching and writing these articles.

3.1 Introducing a federal moratorium on facial recognition

There are currently no enforceable privacy regulations governing the use of facial recognition technology in Canada, nor the collection or retention of FR data [94]. Moreover, uses of FR by police departments and other public institutions are not subject to any enforceable independent oversight [93]. Neither of Canada's primary privacy laws—the Privacy Act, which dictates how the federal government can collect and use personal information, and PIPEDA, its private sector counterpart—mention FR or other biometric data. Despite a joint investigation by four Canadian privacy commissioners deeming Clearview AI's data collection practices illegal, the company has

refused to comply with the privacy commissioners' request that they cease collecting Canadians' data and delete all previously collected data [94].

An immediate solution to this lack of regulation would be to impose a moratorium on the use of facial recognition technology and related data collection practices by all federal agencies, including the RCMP. The federal government could also provide incentives for cities and provinces, especially Quebec, Ontario, and Newfoundland and Labrador, which maintain their own provincial police forces, to enact similar moratoriums.

Moratoriums and outright FR bans have been implemented in many American cities, including San Francisco, Oakland, Portland, Minneapolis, New Orleans, Pittsburgh, and Boston [77]. If instituted in Canada, such a policy would give the federal and provincial governments time to update their privacy laws. It would likewise give researchers and developers time to address many of the problematic issues with FR, such as its propensity towards bias. While such a moratorium would be admittedly limited in scope, covering only public-sector uses of the technology, its effects could still be substantial, as it would legally guarantee the elimination of harm resulting from police use. Research and consultations undertaken during the moratorium could also be applied to the private sector.

In the following chapter, we explore numerous conditions, some technical and some policy-related, for lifting a federal moratorium on FR technology. In other words, we ask: what conditions must be met before the use of FR by public agencies can be considered safe and legitimate?

3.2 Accuracy and bias conditions

A basic condition for the safe use of FR technology by public institutions is that FR models output accurate results and display no bias against protected groups, such as racial minorities or women. An auditing system could be established whereby developers wishing to provide FR technology to law enforcement or other public agencies would need to first demonstrate that their model is

sufficiently accurate and bias-free. However, as we will argue, defining precisely what is meant by “bias" and “fairness" in a technical, legally binding, and universally applicable sense, is both infeasible and ill-advised.

3.2.1 How have researchers defined bias?

We begin with an overview of the many definitions and methodologies researchers have used to detect bias in facial recognition systems. As shown in Chapter 2, researchers have proposed numerous, sometimes contradictory, definitions of fairness. The field of facial recognition has utilized some of these definitions but has also introduced other, more complex methods by which to measure bias.

NIST [29] used false positive and negative rates at a particular system-specific threshold to assess bias. They noted that many FR systems have fixed decision thresholds for all users of the technology.[1] Note that the choice of threshold is directly tied to a system's false positive and negative rates. Increasing the threshold at which a pair's similarity score is flagged as a match decreases the false positive rate (FPR), as the number of false positives decreases, while the number of pairs that belong to the negative class remains unchanged. Similarly, increasing the threshold increases the false negative rate (FNR). The freedom to modify an FR's system threshold to best fit a particular use context (e.g. in policing, where false positive errors may endanger marginalized communities) is therefore highly desirable.

Krishnapriya et al. [43] likewise measured false positive and negative rates at a fixed threshold. They also considered receiver operator characteristic (ROC) curves, which provide a visualization of the trade-off between false positive and false negative rates. More precisely, ROC curves plot the true positive rate ($1 - FNR$) as a function of the false positive rate. The authors explained that comparing the ROC curves of two separate groups (in their case, Black and white subjects) is not an appropriate method by which to detect FR bias, since the groups may have the same FPRs only at very different thresholds. The ROC curves for two of the four FR models the researchers

[1] Here, users refer to those who use the technology (e.g. police departments), not to those subjected to it.

tested showed higher accuracy for Black subjects. However, this result masked the fact that at a fixed threshold, false positive and negative rates differed consistently between the groups across all four models. The researchers examined the distributions of both groups' similarity scores, finding that for Black subjects, the distribution curves of imposter and genuine pairs were shifted toward higher scores relative to the distribution curves of their white counterparts. This divergence was responsible for the discrepancies in the groups' FPRs and FNRs at set thresholds.

Cavazos et al. [11] also utilized both false positive rates at a fixed threshold and ROC curves to measure bias, concluding like Krishnapriya et al. [43] that the latter has the potential to be a misleading gauge of racial bias. Testing four FR models' performance on white and East Asian faces, Cavozos et al. found that the area under ROC curves (known as the AUROC) differed little between demographic groups. However, as in Krishnapriya et al.'s tests, a lower threshold was needed for the white subjects to achieve a given FPR. Moreover, when considering only the extreme left end of the ROC plots—that is, considering the true positive rates associated with very small false positive rates—smaller true positive rates were observed for the East Asian subjects. Since many applications, including policing, frequently demand very small FPRs, this discrepancy has the potential to be significant.

Cavozos et al. also discussed best practices for assessing the overall accuracy of a model using ROC curves. Instead of utilizing all possible different-identity pairs to build an ROC curve, the authors recommend using only pairs from the same demographic groups (same race, gender, etc.), a practice known as 'yoking'. Since demographically different pairs are more likely to receive low similarity scores, their inclusion skews the different-identity distribution leftward, thus artificially boosting the model's TPR relative to its FPR. This, by extension, inflates the model's overall estimated test accuracy.

Cook et al. [19] employed a rather different approach to measuring bias in FR systems, organizing a rally in which 11 commercial FR systems could be tested in real-time. They recruited 363 volunteers, who were instructed to line up to have their photo taken by each system. Immediately

after capturing a subject's image, the FR systems attempted to match it against two datasets containing the subjects' faces, one 'same-day' dataset and one 'historic' dataset, composed of photos dating back four years. For each subject, system-specific similarity scores were calculated between the captured image and potentially matching images flagged by the system. The researchers then estimated overall average demographic effects on system performance by performing linear regression on the average same-day and historic similarity scores for all systems, as well as the average transaction times for each subject across systems. Eleven covariates were considered, including gender, age, race (self-identified), and skin reflectance, which correlates with race. Skin reflectance was found to be a stronger predictor of system performance than self-identified race, with lower reflectance (darker skin) associated with lower system accuracy. Lower reflectance was also associated with longer transaction times.

Phillips et al. [71] studied a possible "other race effect" in East Asian and Western-made FR systems. They theorized that, like humans, who are better able to recognize faces belonging to their own ethnic group—an effect that becomes apparent starting as young as three months old—FR algorithms may be biased in favour of the primary ethnic group of their country of origin. The researchers compared a fusion of five algorithms developed in East Asian countries to a fusion of eight Western algorithms, finding that for East Asian faces, the former fusion algorithm had a superior ROC curve to its Western counterpart. The reverse held true for white faces. They also pitted the fusion algorithms' performance against the performance of real humans. Unlike the humans, who proved to be slightly more adept at recognizing members of their own group, both algorithm fusions had higher AUROCs for white faces than for East Asian faces; however, this advantage for white faces was significantly more pronounced for the Western-made fusion algorithm.

Methods to reduce bias also utilize different definitions of fairness. Terhörst et al.'s [95] method is based on the concept of fairness through awareness, and the researchers reported bias by computing the standard deviation of the different demographic groups' FNRs. Robinson et al. [80] examined detection error trade-off curves (i.e. plots showing FNR vs FPR); their method ensures equal FPRs across groups. Dhar et al. [22] defined bias as the difference between two groups' TPR at

a given FPR, while Gong et al. [28] defined it as the standard deviation between different groups' AUROCs. Wang et al. [102] considered ROC curves and measured each group's TPRs.

Table 1 summarizes the bias definitions and measurement methods used in various papers.

Authors	Bias measurement approach
Grother et al. (NIST) [29]	FPRs and FNRs at fixed thresholds
Krishnapriya et al. [43]	FPRs and FNRs at fixed thresholds, ROC curves
Cavazos et al. [11]	FPRs at fixed thresholds, ROC curves
Cook et al. [19]	Linear regression on average similarity scores and average transaction times
Terhörst et al. [95]	Standard deviation of groups' FNRs
Robinson et al. [80]	Error trade-off curves
Dhar et al. [22]	TPRs at fixed FPR
Gong et al. [28]	Standard deviation of groups' AUROCs
Wang et al. [101] TPRs	ROC curves
Phillips et al. [71]	ROC curves and AUROCs

3.2.2 How should bias and accuracy be measured?

As evidenced by the myriad of definitions employed by different researchers, no universal fairness metric has been established. However, most of the definitions that have been prioritized make use of false positive or false negative rates. One could make the case for the relevance of additional fairness definitions not related to FPRs and FNRs, such as fairness-calibration, that have as of yet been overlooked in the FR literature (for a discussion on the applicability of fairness-calibration to FR, see Chapter 4).

While this lack of a clear, easily testable fairness definition makes designing a simple algorithmic auditing system infeasible, it offers an opportunity for policy-makers to go beyond mathematical one-size-fits-all notions of fairness and truly consider what makes a facial recognition system fair. Fairness and performance are largely context-dependent. Consider three applications of FR: an

office security system using FR to screen arriving workers, a police investigation into a convenience store robbery, where FR technology is used to analyze security camera footage, and a terrorist attack, where live FR is used to catch suspects. In each situation, false positive and negative errors pose different levels of risk to both individuals and the public at large. In the office security case, a false positive error is unlikely to result in great harm to either the intruder or the office workers, while repeated false negative errors, particularly if they affect certain groups of workers at higher rates, could impede workers' perceived job performance and timeliness, thereby impacting their wages and advancement opportunities. Conversely, in the second context, false positive errors engender a substantial amount of risk in comparison to false negative errors. Due to ongoing disparities in police treatment, this risk is heightened if the falsely detained individual is Black or Indigenous, groups disproportionately likely to be mistakenly identified by FR systems. Finally, both false positive and false negative errors pose the highest amount of risk in the third scenario. Because of the high threat level to civilians, individuals incorrectly identified as suspected terrorists are at a considerable risk of police brutality or death. On the other hand, false negative errors could lead to civilian casualties if the perpetrators are not successfully identified.

As the above three scenarios demonstrate, a case-based approach to performance evaluation and bias detection is required. Such an approach necessitates not only a consideration of the most appropriate fairness metric to optimize, but also of which groups deserve particular attention and protection (e.g. Black and Indigenous people), given the algorithm's use case (e.g. law enforcement investigation) and the context in which it will be deployed (e.g. in a police department that may already have a history of racial bias). To help address the first of these considerations, we present in Chapter 4 a simple post-hoc method that would allow FR users to choose which fairness definitions to emphasize.

Any FR auditing system adopted by the Canadian government should likewise take use cases, the environment of deployment, and the most at-risk groups into account. This would involve auditing both the FR developers—analyzing, for instance, the bias mitigation strategies they've employed—and the agencies that utilize their systems. In particular, auditors should investigate

how government agencies use FR in practice and the potential for that use to lead to discrimination. If one police department uses a lower match detection threshold than its neighbours, but the departments all interpret their model's outputs in the same way (i.e. if one department believes that 90% is a sufficiently high probability threshold, while the others only consider potential matches with similarity scores above 97%), then the first department is more likely to make false arrests. Ho et al. [36] suggest using A/B testing to appraise how FR users (e.g. police officers) interpret confidence scores outputted by a model. They also advocate for A/B testing in comparing the decisions users make with or without the help of the FR system and in judging the users' possible over (or under) reliance on the technology.

The EU's proposed AI Regulation is an example of a regulatory system that prioritizes the process by which an algorithm is developed and deployed over one specific performance or bias threshold that the end product must meet [45]. Indeed, Watcher et al. [100] argued that legal notions of fairness, at least in the EU, are too context-dependent and subject to judicial interpretation to ever be fully "automated".

Ultimately, when shaping policy or designing auditing systems to ensure that the FR systems used by government agencies are fair and unbiased, it is imperative that we understand that bias stemming from the FR model itself can be greatly compounded by the socio-technical environment in which the model is used.

3.2.3 What testing dataset should be used?

As described above, a national moratorium should only be lifted once a robust auditing mechanism is established to verify FR systems' propensity toward bias. This auditing process would require analyzing algorithms' performances on testing datasets. Crucially, these testing datasets should not be publicly available, or else developers could train their models on them.

The choice of testing dataset is paramount: while the testing dataset cannot impact an FR algorithm's underlying accuracy, it can certainly alter the algorithm's perceived accuracy. For example,

testing a racially biased algorithm on a dataset that is largely racially homogenous would yield a better overall accuracy than testing the same algorithm on a balanced dataset, thereby obscuring the model's true performance. Many commonly used testing datasets skew white. 80% of IJB-A's images depict light-skinned individuals; for Audience, that percentage reaches 86% [8].
Cavazos et al. [11] discussed the possible impacts of testing FR performance on datasets that do not adequately represent the population on which the algorithm will operate. To rectify this problem, many researchers have introduced demographically balanced datasets. The methods they employed to produce these datasets could help inform federal auditors when creating their own datasets.

Robinson et al. [80] constructed Balanced Faces in the Wild (BFW), a racially and gender balanced dataset containing 20,000 images of 800 individuals. BFW was created by sampling subjects from VGGFace2, a 3.3 million image dataset. The researchers used a pre-trained ethnicity model to scan VGGFace2's list of names, then later used gender and ethnicity classifiers on the selected names' corresponding images to narrow down their pool of potential subjects. Out of this pool, they manually selected 100 individuals from each of their eight demographic groups—Asian females, Asian males, Black females, Black males, Indian females, Indian males, white females, and white males—to add to BFW. Once their final list of subjects was solidified, they used MTCNN to identify all associated facial images. Using the FR model Sphereface, they generated face embeddings and built a matrix of similarity scores for each individual's face images. Faces that proved too dissimilar to their subject's other images—that is, faces with low median similarity scores—were removed from the dataset[2]. The remaining images were visually verified.

Wang et al. [102] introduced Racial Faces in-the-Wild (RFW), a testing dataset which the authors used to measure the racial bias of eight FR algorithms. RFW is comprised of 12,000 identities and 40,000 images, equally split between four racial groups: Asians, Indians, Africans, and Cau-

[2]This was done to purge possible false positive pairs from the dataset. We note, however, that since many individuals do drastically change their appearance over time, such a removal practice risks making the dataset less representative of the real images that FR systems may encounter and less useful for testing an FR system's performance on "difficult" pairs.

casians. While subjects belonging to the former two groups could be directly collected from the MS-Celeb-1M dataset using its “nationality" attribute, identifying subjects belonging to the latter two groups required applying an ethnicity classifier, Face++, to the dataset and manually verifying selected images that were assigned low confidence scores. Individuals were also excluded if some of their associated images were predicted to be of different races or if their nearest neighbour in two other datasets, CASIA-Webface and VGGFace2, as determined using an Arcface FR loss, shared their identity. The researchers manually cleaned all 40,000 images; they also verified that the pose, age, and gender makeups of each racial group were similar.

Alvi et al. [2] built Labeled Ancestral Origin Faces in the Wild (LAOFIW), which contains 14,000 images of individuals of various ancestral origins, namely sub-Saharan Africa, the Indian Subcontinent, Europe, and East Asia. These images were selected using the Bing Image Search API. The dataset contains a diversity of poses, illumination levels, and facial expressions, and is split approximately evenly between male and female subjects.

All three of the above datasets are demographically balanced between four distinct racial groups. However, as Cavazos et al. [11] pointed out, race cannot be neatly divided into discrete categories. In testing an FR algorithm only on photos of white, Black, Indian, and East Asian individuals, bias against mixed-race subjects and against subjects belonging to other racial groups, such as Indigenous North Americans, cannot be evaluated. This is a grave oversight, as Indigenous people already face disproportionate police scrutiny in Canada, a trend that could be amplified by biased FR systems.

Bias could instead be proposed to be measured as a function of skin reflectance and not race, as Cook et al. [19] found the former to be more strongly related to FR performance. Unfortunately, this solution would likely be inadequate, as different racial groups, such as whites and East Asians, are apt to be lumped together, thus obscuring potential bias against certain groups.

Additionally, one key component of the “context" consideration for testing both bias and system accuracy is the specific real-world conditions in which a model is deployed. For example, a model

may be trained on high-quality images but later be required to process low-quality CCTV footage. As we will see in a later section, this may influence the degree of bias manifested by the system, in addition to its overall accuracy level. Testing only on high-quality images may misrepresent the model's true accuracy and propensity toward bias. However, to complicate matters, there may be ethical dimensions related to consent and privacy of testing on "realistic" data like CCTV images (see section 3.3.2 for a further discussion of the ethics of data collection for testing purposes).

3.2.4 Which demographic groups should be protected?

While we have highlighted the need to prioritize at-risk groups when measuring bias in different use cases, it also crucial to extend our discussion to include groups that are often not considered when analyzing FR technology. So far, our discussion has focused on racial bias and to a lesser extent bias related to age and sex (male vs female). Indeed, the vast majority of research and legislative efforts have centred on these demographic factors. The Canadian Charter of Rights and Freedoms precludes discrimination on the basis of race, age, and sex, but also on the basis of other attributes, including physical and mental disability [56]. Little has been done to investigate the performance of FR on subjects with craniofacial differences such as cleft lip, Pierre Robin sequence, or Treacher Collins syndrome [9]. Facial recognition tools have been used to help detect genetic conditions like Down Syndrome [74] (a task which itself, in a 2017 study, yielded far better results for white Belgian faces than Black Congolese ones [46]), but individuals with these conditions have not been considered when testing standard FR systems for bias. People with facial disfigurements form a sizeable group: in 2017, they accounted for nearly 1% of the total population of the UK [92].

For a moratium to be lifted, it is vital that auditors be able to assess FR bias against Canadians with disabilities. This will require identifying the most effective ways to detect such bias. As facial characteristics vary broadly between conditions and syndromes, and even between individuals with the same condition (particularly between individuals of different races), evaluating for bias is not as simple as inserting a small number of images into a testing set. Since this area remains

largely neglected, significant research is required to study and, if necessary, combat, disability and disfigurement-related FR bias. Particular attention should be paid to individuals who have undergone significant facial surgery, as current FR systems are unlikely to be able to match new and old photos of these individuals. Before removing FR restrictions, the federal government should implement a manual oversight and/or appeals process to ensure that such individuals are not discriminated against, for instance by FR systems designed to detect fraud in passport applications.

Transgender, non-binary, and other gender non-conforming individuals constitute another group that has been excluded from studies on FR bias. Keyes [39] found that nearly all papers on automatic gender recognition papers define gender as binary, and a large majority define it as immutable. The same appears to be true for facial recognition papers. Many popular commercial gender recognition algorithms, which operate very similarly to facial recognition algorithms, have been shown to misclassify transgender individuals at high rates and to be totally unable to classify non-binary individuals [88].

In an attempt to mitigate this problem, Wu et al. [105] included training images of non-binary people, as well as a non-binary prediction option, in their gender recognition algorithm. However, Keyes [40] criticized Wu et al.'s paper for its choice of dataset (Wikipedia photos of non-binary celebrities), arguing it did not accurately represent non-binary people as a group. They further objected to the paper's use of 'non-binary' as a discrete third gender category, explaining that the physical appearance and outward gender expression of non-binary people differ greatly between individuals and cannot be essentialized as simply "androgynous". While the aforementioned works did not directly examine potential bias against trans and non-binary people in facial recognition applications, they did highlight the complexity of both addressing such bias and of building an appropriate testing set to measure it.

There have seemingly been no studies comparing the performance of FR systems on trans men and women and non-binary people to their performance on cisgender men and women. As discrimina-

tion based on gender identity or gender expression is prohibited in Canada, it would be misguided to deem any FR system “safe" and “unbiased" before any such studies are undertaken.

Like with individuals who have had substantial surgery because of a facial disfigurement, there is also a risk that certain FR applications (e.g. fraud detection) may lead to discrimination against trans people. While limited efforts have been made to improve FR algorithms' ability to match photos of the same transgender individual at different points in time [44], it is nevertheless imperative that government agencies wishing to employ such uses of FR guarantee effective recourse mechanisms to handle errors.

Most crucial of all is the question of intersectionality: how do we measure bias against individuals belonging to more than one at-risk group? Kearns et al. [38] explored the “fairness gerrymandering effect", whereby a model is unbiased on the basis of individual attributes (e.g. race and gender) but exhibits bias when tested on intersectional subgroups (e.g. Black women, white women, Black men, and white men). They introduced an algorithm to find the classifier that is maximally fair with respect to a set of intersectional subgroups; they defined fairness as requiring low false positive (or alternatively, false negative) rates across all groups. We note, however, that such a solution requires discrete demographic groups, which as previously established, do not exist.

Detecting bias along intersecting demographic axes is key to ensuring that models are indeed bias-free. As previously stated, it is likewise key to recognize that in certain use cases some groups are at a higher risk of harm than others: bias against a white female has different implications in the context of law enforcement than bias against a Black male. A context-based auditing system must therefore possess the ability to analyze the deployment context in order identify at-risk groups, as well as the ability to detect bias, both model-based and environmental (e.g. differences in how FR users interpret the model's output when different groups are concerned).

3.2.5 What causes bias?

As stated in [67], “for a moratorium to be lifted, it is not enough to possess the means by which to test bias—FR systems must be unbiased enough to pass said test". Thus, eliminating FR bias is an implicit condition for lifting a federal moratorium on the technology. Unearthing the root causes of that bias is a key starting point for eradicating it.

According to Cavazos et al. [11], one phenomenon explains “both everything and nothing" about demographic bias: the discrepancy between different groups’ similarity score distributions. In Figure 3.1, we show the distribution of three FR models’ cosine similarity scores on four racial groups.

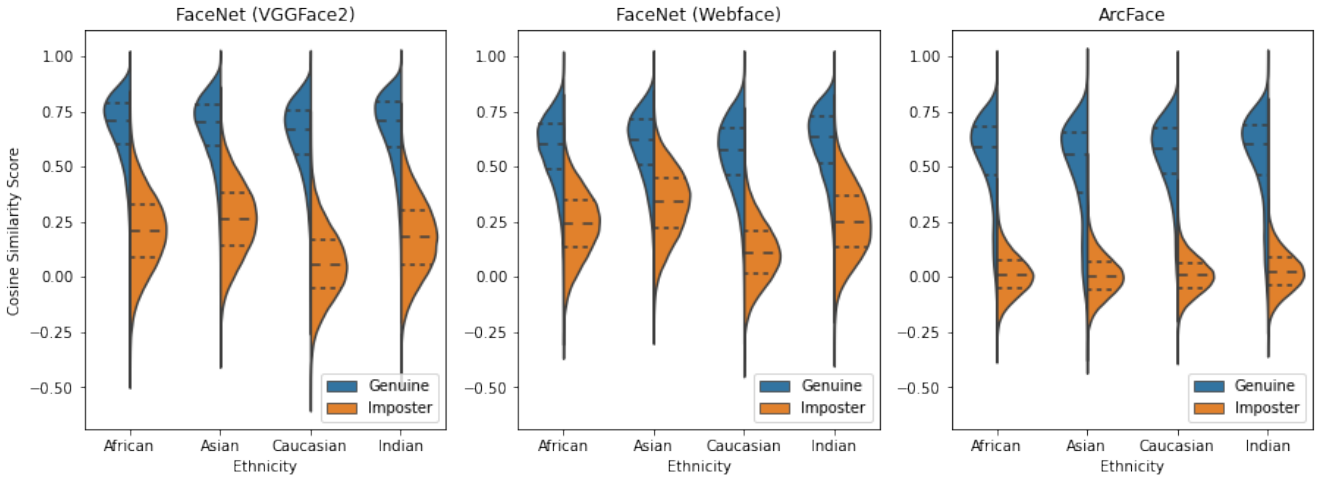

Figure 3.1: Distribution of cosine similarity scores by ethnicity on the Balanced Faces in the Wild dataset. Figure source: [83].

Cavazos et al.’s explanation is helpful only if the cause (or causes) of these diverging distributions can be unearthed. Researchers have presented several potential culprits, but measuring the influence of these factors on a large and complex model’s level of bias is highly difficult.

Image quality

Krishnapriya et al. [43] examined the MORPH dataset, a large collection of mugshots, and observed that images of Black subjects tended to be of lower quality than those of white subjects:

only 48% of Black images met the International Civil Aviation Organization's (ICAO) standards for travel documents compared to 57% of white images. They noted that the two groups' images especially differed on the basis of "brightness" and speculated that the quality gap could be closed by adopting skin tone-appropriate lighting. However, in analyzing only the subset of the groups' photos that were ICAO-compliant, they found a slightly smaller gap between the groups' ROC curves compared the full dataset case, but little difference in their FNRs at a low FPR.

Cavazos et al. [11] attempted to measure racial bias as a function of image difficulty. They compared the ROC curves of four FR algorithms on the Good, Bad, and Ugly (GBU) dataset, which divides its face pairs into three subsets based on how easy or how difficult it is for an FR algorithm to accurately match them. This partition is made by computing similarity scores for each pair using top-ranked FR systems (as judged by NIST's 2006 FR vendor test). They found that for "good" images, all four models achieved close to perfect accuracy on both white and East Asian faces, though for two of the four, a small bias against East Asian faces was detected at very low FPR rates. For the "ugly" set, three of the four models had starkly superior ROC curves for the white images.

NIST's Grother et al. [29], in their updated FR Vendor Test, remarked that increased image quality leads to smaller discrepancies between various groups' false *negative* rates. However, their results showed that inflated false *positive* rates for certain groups were still apparent even for high-quality photos. Thus, it is unlikely that the problem of demographic bias in facial recognition can be solved solely by ensuring that equally high-quality photos are taken for all demographics. Moreover, even if such a solution were sufficient, it would be difficult to maintain in practice, particularly in the context of police surveillance and border crossing.

Balanced training set

Cavazos et al. [11] discussed the possible impact of a demographically unbalanced training set on bias but concluded that assessing such an impact is troublesome, as many additional factors are at play, such as image quality.

Phillips et al. [71] considered an “other-race effect" in Western and East Asian FR systems. Such an effect, they speculated, could be attributed, at least in part, to differences in the racial makeup of the algorithms’ training sets. Published in 2011, the numerical results of this work are out of date, but its central question of whether the country of origin of an algorithm affects its propensity toward bias remains highly relevant. In their work for NIST, Grother et al. [29] likewise noted that while East Asian faces had high false positive rates across most FR algorithms, this was not the case for Chinese-made systems; indeed, for these systems, East Asian faces had notably low FPRs.

Krishnapriya et al. [43] noted that ResNet’s superior ROC curve for Black subjects compared to white subjects was achieved despite the fact that the model’s training set skewed heavily white. They claimed that this implied that equal accuracy between groups could be attained without a demographically balanced training set. However, the authors remarked elsewhere in the paper that ROC curves are a poor measure of bias. Moreover, even if true, this claim would not preclude a balanced training set from positively impacting a model’s level of fairness.

3.2.6 How can bias be reduced?

Drawing on techniques from various other fields of machine learning, researchers have devised a number of strategies to reduce bias in facial recognition algorithms. These strategies include employing adversarial training [2, 22, 28], reinforcement learning [101], data augmentation [42], domain adaptation [102], group-specific thresholds [80, 95], and post-hoc neural networks [96]. All of these methods, with the notable exception of group-specific thresholds and post-hoc neural networks, require retraining the FR model, which may be computationally expensive or even infeasible in practice.

Adversarial training

Dhar et al. [22] introduced an ‘Adversarial Gender De-biasing algorithm’ (AGENDA), with the goal of curbing the gender predictability of face embeddings. This strategy hinges on the assumption that biased demographic information gets transferred from the dataset to the face embeddings

while training for identity classification and that removing this demographic information would reduce gender bias in facial recognition tasks. AGENDA aims to transform face embeddings so that they cannot be used to predict gender. The algorithm consists of four principal components. The first is a pre-trained network P, which outputs a face embedding $f_{in} = P(I)$ given an image I. The other components are a generator model M, which takes in f_{in} and generates a lower dimensional embedding f_{out}; an identity classifier C, which takes in f_{in} and generates a prediction vector for identity classification; and an ensemble of K gender prediction models $E_1, E_2...E_K$.

The goal is to train M so that it generates face embeddings that C is capable of identifying but that contain limited gender information and therefore fool the gender discriminator E. An E that is totally gender-agnostic (i.e. one that produce posterior probabilities of 0.5 for both groups of face embeddings - male and female) implies that the face embeddings contain no gendered information whatsoever. For additional details, see Appendix B.1.

Dhar et al. defined gender bias as the difference (in absolute value) between the two groups' TPRs at a fixed FPR. They found that applying AGENDA to face embeddings reduced gender bias for two test datasets. This reduction came at the expense of slightly worse to significantly worse TPRs at different fixed FPRs.

Gong et al. [28] also introduced an adversarial network, which they dubbed DebFace. The network contains an identity classifier, DebFace-ID, and three demographic classifiers, DebFace-G, DebFace-A, and DebFace-R, which are trained to predict gender, age, and race. They argued that a single-demographic adversarial network is inadequate, as a gender prediction model can, for example, still display racial bias. They instead proposed a novel multi-task framework consisting of four classifiers, jointly trained via adversarial learning to be able to correctly predict their own attribute (gender, age, race, or identity) while being agnostic toward the other three.

Gong et al. tested DebFace on two genders, six age groups, and four races (a total of 48 demographic groups). They defined bias as the standard deviation of the groups' AUROCs. Compared to face embeddings generated by the shared feature encoder of DebFace (i.e. the embeddings gen-

erated prior to the adversarial training stage), the DebFace-ID embeddings were found to contain less gender-, age-, and race-based bias when used to perform facial recognition. They likewise found that DebFace was less biased when performing gender, age, and race prediction than the pre-adversarial model (which the researchers called BaseFace). However, BaseFace slightly outperformed DebFace-ID at facial recognition on all four test datasets, since removing demographic information limits the classifier's ability to identify faces.

To address this decrease in accuracy, Gong et al. combined the four DebFace embeddings into an aggregate embedding, DemoID. DemoID accrued additional bias but also additional accuracy in comparison to DebFace-ID, allowing it to outperform BaseFace on the test datasets. In other words, DemoID struck a trade-off between bias and accuracy: it proved more biased than DebFace-ID but less biased than BaseFace and proved more accurate than both models.

The authors noted that their solution failed to fully exterminate bias: DebFace still performed differently on different demographic groups. They attributed this to disparities in image quality or capture conditions between the groups.

Alvi et al. [2] did not utilize the term "adversarial learning" in their work, but the method they introduced bears a strong resemblance to those described above. Their aim was to create a network capable of learning a single classification task (e.g. age recognition) but that is blind to other information (e.g. pose, ethnicity, and gender). Their algorithm, which they termed a joint learning and unlearning algorithm, involves the use of a standard convolutional neural network[3] to output a face embedding. The face embedding is passed to a primary branch, consisting of a primary classification loss, and to multiple secondary branches. Each secondary branch is associated with a different variable (e.g. gender) and contains both a classification loss and a confusion loss. For the case of gender information, the former loss learns to predict the associated gender of the face embedding, while the latter modifies the face embedding so that it becomes invariant to shifts in

[3]For more information on convolutional neural networks, see [82].

gender. Training occurs in two alternating stages: the secondary classification losses are trained together, followed by the primary classification loss and the confusion losses.

While the authors didn't utilize their method to debias face recognition, limiting their scope to age, ethnicity, gender, and pose classification, their work could easily be applied to such a task.

Alvi et al. found that their joint learning and unlearning algorithm could remove between 78% (removing ethnicity information from a pose classifier) and 100% (removing ethnicity information from an age classifier) of secondary information from a primary classifier. Here, the percentage of unlearned information corresponds to the relative difference between the model's baseline accuracy and its debiased accuracy compared to random chance. The researchers also considered the impact of their method on overall accuracy: while the age classifier's accuracy decreased by 5% post debiasing, no change in accuracy was noted for the gender and pose classifiers, and the ethnicity classifier's accuracy actually improved by 1%.

Reinforcement learning

Wang and Deng [101] proposed a novel method, a reinforcement learning based race balance network (RL-RBN)[4]. Standard FR models maximize their overall accuracy by applying margins between different classes (e.g. dogs and cats in image recognition). In the case of FR, each class consists of one individual subject. If a certain type of subject (e.g. white people) are highly represented in the training set, the model will optimize for that group, to the detriment of other groups, as that will help improve its overall performance. RL-RBN hinges on the idea of modifying the margins between subjects in order to achieve a balanced accuracy across different racial groups. The authors noted that the same process could be applied to mitigate age or gender-based bias. See Appendix B.2 for a more detailed explanation of RL-RBN.

Wang and Deng tested their method by training their model, as well as three other state-of-the-art models, on two custom-made datasets, BUPT-Globalface and BUPT-Balancedface, and evaluating

[4]For more information on reinforcement learning, see [53]

their respective performances on the Racial Faces in-the-Wild dataset. They found that their model obtained a more balanced performance, and in some cases higher accuracies for each of the four racial groups, than other models.

Data augmentation

Kortylewski et al. [42] tackled bias caused by variations in face poses by inserting synthetic images into the training set. While their method doesn't directly address demographic bias, it could be readily modified to instead consider race, gender, or age-based bias. The authors undertook two major experiments: firstly, they used labelled synthetic data to measure an FR model's overall accuracy as a function of face pose. While there already exists many demographically labeled datasets, an image generator like the one proposed by Kortylewski et al. could create images with fine-grained annotations that specify, for instance, the precise age or skin tone of the synthetic subject.

The authors' second experiment involved training a model with a mix of genuine images and a large number of synthetic images that greatly varied in face pose. The benefits of this approach were twofold: when testing on genuine images, the synthetically trained model was better able to generalize to new face poses, and it was found that the number of genuine images needed for training could be slashed by 75% without significantly affecting the model's overall accuracy. The ability of such an approach to reduce demographic bias could and should be investigated.

Yin et al. [107] introduced a method by which to augment the feature space[5] of subjects who only have a few images in the training set. Their method, Feature Transfer Learning (FTL), generates new images for these underrepresented samples by mimicking the feature distributions of regular subjects. This requires assuming that the features of both regular and underrepresented subjects are normally distributed and share the same intra-subject variance. Like Kortylewski et al.'s [42] work, this method is not directly applicable to the problem of FR bias. However, a similar approach

[5]The high-dimensional space in which "features", i.e. pieces of information about the images, exist.

could be envisioned to generate entirely new subjects and sets of associated images by replicating the feature distributions of subjects in the same demographic groups.

Domain adaptation

Wang et al. [102] designed a deep information maximization adaptation network (IMAN) based on unsupervised domain adaptation (UDA), a method that has been shown to decrease algorithmic bias. They found that IMAN achieved higher true positive rates than other FR models on Indian, Asian, and African subjects from the authors' newly introduced Racial Faces in-the-Wild dataset (see section 3.2.5). They likewise compared different models' ROC curves, with IMAN coming out on top. A full breakdown of their method can be found in Appendix B.3.

Group-specific thresholds

Having observed differences between various demographic groups' FPRs when employing a single threshold for all subjects, Robinson et al. [80] introduced group-specific thresholds. Using three FR models, they tested their method on Labeled Faces in the Wild (LFW), a dataset in which eight groups—images of Asian, Black, Indian, and white females and males—are labeled. The group-specific thresholds, which ensured that each group had the same FPR, were determined using detection error trade-off (DET) curves, graphical plots that map a group's FNR as a function of its FPR. Optimal thresholds were found to differ between FR models.

During testing, one image from each pair was designated the 'query' and the other the 'test'. The threshold associated with the query's demographic group was adopted for the pair. The authors measured the groups' TPRs at different fixed FPRs, noting improvements across all groups relative to the global threshold approach.

In practice, demographic labels are not always available. Efforts to label search databases may themselves be prone to bias (either human or algorithmic) and may breach subjects' privacy. Moreover, many attributes, such as race, resist categorical labelling. It is thus desirable to devise a

method that assigns different thresholds to different individuals even when group labels are not known.

Terhörst et al. [95] did just that, introducing a fair score normalization (FSN) method that splits the dataset into clusters and assigns to each cluster a unique threshold. Their work was also motivated by their assertion that many previous bias mitigation methods, such as adversarial networks, are computationally expensive and difficult to integrate into existing FR systems. Moreover, they noted that these methods typically curb the model's overall accuracy.

Their fair score normalization method is based on the notion of individual fairness, or fairness through awareness (see Appendix A.3.3), which requires that similar individuals be treated similarly. It involves two phases: a training phase and an operation phase.

The training phase proceeds as follows: given a set of face images split into training and testing sets, face embeddings are computed by an FR model. A k-means cluster algorithm is then trained to divide the training set into k clusters. Using each cluster's genuine and imposter scores, cluster-specific thresholds are calculated to ensure a universal FPR of 10^{-3} (as recommended by Frontex, a European border guard agency).

The operation phase involves computing a normalized similarity score $\hat{s}_{i,j}$ for each pair of images i and j in the testing set. For the images in a pair, their clusters and associated thresholds thr_i and thr_j are identified. These thresholds, along with a global threshold thr_G and the pair's original similarity score s_{ij} as outputted by the model, are used to determine

$$\hat{s}_{i,j} = s_{ij} - \frac{1}{2}\left(\Delta thr_i + \Delta thr_j\right) \tag{3.1}$$

where

$$\Delta thr_i = thr_i - thr_G \tag{3.2}$$

Terhörst et al. tested their approach on three datasets using two FR models: FaceNet and VGGFace. They fixed an FPR of 10^{-3} for all clusters and used FNR as their performance metric. For each dataset, they considered up to three demographic attributes: age, gender, and ethnicity, depending on which labels were available. They found that their method decreased FNRs for most demographic groups, with the exception of already exceptionally high-performing groups like white or middle-aged subjects in the ColorFeret dataset, whose FNRs increased after their scores were normalized. The researchers measured a system's level of bias for a particular attribute (i.e. age group) by calculating the standard deviation of the demographic groups' FNRs. They found that their method reduced total bias by 82.7%. Overall performance (as measured by the overall FNR) increased by 82.9%.

Investigations into the optimal value of k for the k-means cluster algorithm revealed that both small and large k values lead to reductions in performance. For small k values, this was due to clusters of unequal sizes, while clusters containing too few images to produce reliable thresholds were responsible for the poor performances associated with large k values. A value of $k = 100$ was found to be optimal.

Figure 3.2 shows a visualization of the Audience dataset's FaceNet embeddings, where each point corresponds to an individual.

Post-hoc neural network

Terhörst et al. [96] introduced a fair template comparison (FTC) method requiring no re-training of an FR model. Given a model capable of generating face embeddings, instead of simply comparing two embeddings' cosine similarity scores, they passed the embeddings through a neural network trained to output fair scores. See Appendix B.4 for details.

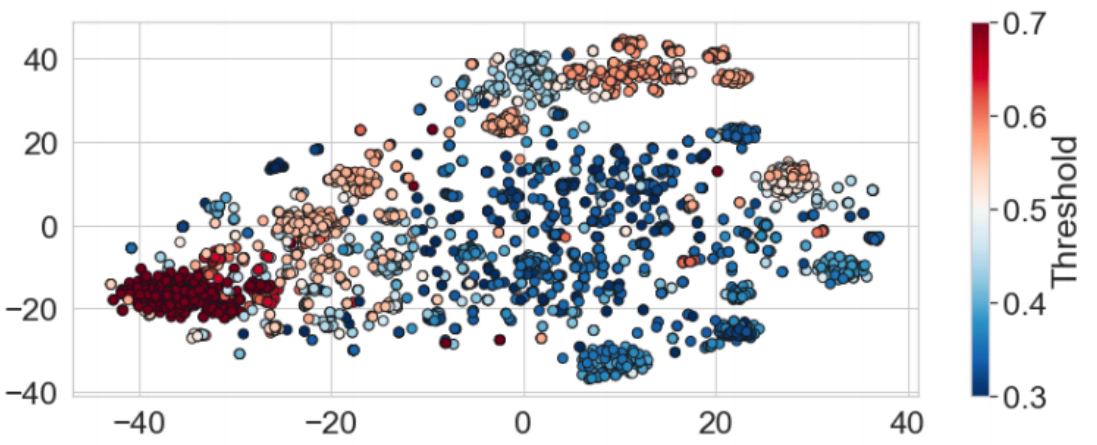

Figure 3.2: Audience dataset's FaceNet embeddings. Each point corresponds to an individual and is plotted using a t-SNE algorithm, which maps high-dimensional embeddings to two dimensions. Different colours correspond to different optimal thresholds. Clusters of similar individuals requiring similar thresholds are readily apparent. Figure source: [95]

Terhörst et al. tested their method on the ColorFeret and Labelled Faces in the Wild datasets and measured bias as the Mean Absolute Deviation of groups' TPRs at a fixed global FPR. They found that their method reduced bias by up to 52.67%.

Could the bias problem be solved simply by substantially improving the accuracy of FR models?

Studies measuring the bias present in various FR systems have observed significant differences between the best- and worst-performing systems. Cook et al. [19] found that systems with high overall performances fared better on images with low reflectance (and by extension, dark skin) than inferior systems did on all photos.

Similarly, while NIST's Face Recognition Vendor Test detected bias across a majority of systems, the highest-performing one-to-one verification systems showed little bias, with consistently low FPRs and FNRs across demographic groups [29]. The best one-to-many identification systems exhibited even less bias. In some systems, such as NEC-3 and Idemia, the disparities between groups' FPRs were deemed "undetectable", while many of the most accurate systems actually had

lower FNRs for white women than for white men (Black men had lower FNRs than white men across most systems) [29]. Moreover, even when disparities between groups were present, some systems performed so well on even the most “difficult" groups that these disparities were virtually meaningless. For example, an algorithm developed by the American firm Calvi exhibited an FPR that was 13 times higher for Indigenous women than for white men; however, their system was so effective that Indigenous women were still only false identified at a rate of around 0.01% [48].

This suggests that the problem of bias may naturally disappear as the overall accuracy of FR algorithms continues to increase. This would of course be a welcome development, but there are numerous reasons for researchers and policy-makers to continue to engage with the issue of bias, certainly for the foreseeable future if not for longer. Firstly, as demonstrated by NIST, a large number of commercially available FR system are in fact biased against certain demographic groups. Many other systems have not been tested by NIST, including Clearview AI’s; in fact, the company has failed to self-report any information regarding bias and has self-reported wildly inconsistent estimates of its overall accuracy, ranging from 75% to 98.6% accurate [32]. Even if it is indeed possible to mitigate bias simply by virtue of achieving a high overall accuracy, it remains crucial that FR systems are audited to ensure that they do in fact display negligible bias. Secondly, as discussed above, bias may stem not only from the FR algorithm but also from the environment and the way in which it is used. For example, NIST found that six of the best-performing systems had higher FPRs on Black men than white men at a certain threshold, but that the opposite was true at another threshold [48]. Audits must consider an FR’s performance not in isolation but as a part of a larger socio-technical system.

3.2.7 Checklist: what needs to be accomplished before a moratorium can be lifted?

Policy challenge

- Build an auditing system to measure bias in FR models used by government agencies. The auditing system should scrutinize developers, users, and the model itself. It should consider context to determine the best fairness definition(s) and to analyze the socio-technical environment in which model is deployed (which demographic groups are most at risk? How does the agency use the FR system?).

Research challenges

- Develop effective context-dependent methods by which to mitigate bias.
- Study the implications of different fairness definitions to identify which definitions are most relevant for different use cases.
- Test FR systems on neglected groups such as people with facial disfigurements and transgender people.
- Study how to best measure bias when discrete demographic groups do not exist (e.g. racial categories).
- Study the technical and legal considerations of prioritizing measuring bias against particular groups (e.g. Black men in the context of policing).

3.3 Data conditions

Data usage is a dimension that extends beyond bias. As we have previously discussed, bias is heavily tied to data: the choice of which training and testing datasets to employ can affect a model's

underlying bias levels, as well as whether and how that bias is detected. Choices around data also have major implications for privacy and consent. Often, these two issues are intertwined, as bias mitigation and measurement usually requires knowledge of sensitive demographic information. The new EU AI Regulation accounts for this, specifying that FR providers may process special types of personal data, such as those specified by Article 9 of the EU's General Data Protection Regulation (e.g. data relating to an individual's race, sexual orientation, etc.) [18], for the purpose of bias detection; however, as-of-yet unspecified safeguards must be in place to protect individuals' privacy and other rights [14].

Unlike the EU, Canada has no law nor proposed law regulating AI, and its privacy and data protection laws are woefully out-of-date. Before a Canadian moratorium on FR could be lifted, the government would need to complete the following two steps: update its laws, including the Privacy Act and PIPEDA, and develop a framework and complementary auditing system to ensure that neither FR providers' nor users' data practices contravene the fundamental rights of Canadian residents. Below, we briefly outline the current state of data protection laws in Canada and their ambiguous relationship with FR technology. We then highlight some key considerations for the development of a FR data governance framework and auditing system.

3.3.1 Data and privacy laws

The two primary laws regulating personal data in Canada are the Privacy Act, which governs its collection and use by the Canadian government, and the Personal Information Protection and Electronic Documents Act (PIPEDA), which governs its collection and use by private companies for commercial purposes [59]. The Privacy Act applies only at the federal level: provincial governments are regulated by provincial laws [59]. Similarly, PIPEDA does not apply to data that remains within its province of origin; companies that operate solely within Alberta, British Columbia, or Quebec are instead subject to provincial laws [59].

The ability of the Privacy Act or PIPEDA to adequately regulate FR data is severely limited. Unlike the EU's General Data Protection Regulation (GDPR), the UK's Data Protection Act, and California's Consumer Protection Privacy Act, neither of the Canadian laws explicitly mentions biometric data (of which facial recognition data is an important type) [93]. Thus, despite its extremely sensitive nature, FR data is not subject to any additional rules. Even enforcing the laws as they currently stand has proven difficult: the Office of the Privacy Commissioner has argued that Clearview AI violated PIPEDA during its mass web-scraping operations; however, its condemnation of the company carries no legal weight. Clearview AI, for its part, has maintained that it is not subject to PIPEDA, claiming it has "no substantial link to Canada" [58].

Stevens and Brandusescu [93] argued that Canadian laws "neither explicitly allow nor prohibit law enforcement agencies from processing biometric information." Notably, their investigation failed to identify any Canadian laws, at the provincial or federal level, that restrict law enforcement agencies' ability to acquire FR technology or to process FR data. Stevens and Brandusescu further asserted that due to the Privacy Act's relatively lenient approach to consent, federal agencies can provide personal data (and thus FR data) to the RCMP, conditional only on the latter specifying in writing which data they require and stating that the data will be used for policing.

The Canadian government introduced Bill C-11 in November 2020 with the intent of modernizing Canada's data laws. Numerous experts have criticized the bill, including Canada's Privacy Commissioner Daniel Therrien [5], who explicitly condemned its laxity toward FR.

It's clear that current and currently proposed Canadian data legislation is ill-equipped to respond to the threats posed by FR. This inadequacy is further reason to impose a federal moratorium, which would give the government time to update Canada's laws and to define and implement a governance framework that would establish transparent standards in the collection, disclosure, and use of FR data, and an auditing system to ensure compliance.

3.3.2 Key considerations for a data usage framework

Types of data

How should policy-makers approach designing a data framework? First, we must note that FR systems incorporate four types of data: training data, which may be used to either train the model or to validate its results; external testing data; input data, i.e. data processed by the model after deployment (e.g. images captured by security cameras); and search data, that is, data contained in a search database (e.g. mugshots that the FR system will comb through to identify possible matches). Additionally, FR systems generate data—typically similarity scores for particular pairs of images.

The first four types of data can originate from different sources: a facial recognition system might be trained on images of celebrities [6] and then applied to a police database of mugshots. Search data may be held by the FR users (e.g. mugshots) or may be provided by the FR developers (as in the case of Clearview AI's web-scraped database). Each data category likewise invokes slightly different ethical considerations; for example, there is less risk—both of material harm and of privacy violations—associated with including an individual's image in a training dataset accessed only by developers than with including it in a police search database or with inputting it into an FR model. Conversely, while it is important to ensure that training and testing data is demographically balanced in order to mitigate model bias, this is of course not a requirement for input and search data. The different datasets are also likely to contain different types of personal information: for bias mitigation purposes, training data may have demographic labels, while testing data may likewise have such labels in order to measure that bias. Search databases, by construction, contain names or other personally identifiable information; depending on the context, they may contain demographic labels as well. It is thus crucial that when creating a data governance framework policy-makers do not treat FR data as if they are homogeneous.

[6] Note that such a dataset may not adequately capture the full range of human faces, as images of celebrities are often photoshopped, well-lit, and depict makeup-clad individuals who conform to Eurocentric standards of beauty.

Existing data principles

Next, any data framework and auditing system should be built on existing, albeit insufficient, data conventions, such as those outlined in PIPEDA. The private sector privacy law lays out 10 key principles, namely company accountability; identification of purpose prior to data collection; consent; limiting collection to only necessary data; limiting use, disclosure, and retention so as to fulfill a pre-determined purpose and no more; accurate and up-to-date personal information; security safeguards; transparency about the company's practices; the right to be informed when and for what purpose one's information has been collected and to amend it if necessary; and the right to challenge a company's compliance with PIPEDA [62]. Many of these principles fit under the umbrellas of "purpose limitation", which holds that data should only be collected to fulfill a pre-stated purpose, and "data minimization", which holds that the data collected should be relevant and not disproportionate to the intended purpose, and that it should be retained for no longer than is necessary and shared no more than is necessary [24, 37]. For instance, the Swedish Data Privacy Authority concluded that applying FR to school children to track attendance violated GDPR's purpose limitation and data minimization principles, as outlined in Article 5 [17, 24]. Note, however, that these same principles are often difficult to satisfy in the context of FR, and indeed in machine learning applications more broadly: policy-makers must pay special heed to ensure they can be guaranteed.

Data collection

When drafting the data framework, policy-makers should consider the following questions: what types of data should be allowed to be collected for the purpose of facial recognition by a government agency? Who should be permitted to collect it: the agency, the FR developers, or a third party? How much data should be allowed to be collected? For how long should it be retained? The data minimization principle dictates that the amount of data collected and its retention time should be limited to no more than is necessary to achieve one particular purpose. FR presents numerous complications, however: data collected for the purpose of training, testing, or using FR, along with data generated by the model, may be extremely valuable for further improving the FR

system or reducing its bias. Moreover, limiting data retention may be complicated by the desire for FR providers to maintain long-term search databases that they can offer to different police departments. Despite this, AI Now and UK's Information Commissioner's Office both argued that data minimization is nevertheless a crucial condition to maintain: just because data might be useful for a later purpose does not make it *necessary* to fulfill that purpose, "nor does it justify its collection, use, or retention" [37].

According to legal experts, Bill C-11's commitment to data minimization is questionable: its minimization requirements apply only to the collection, and not to the usage or disclosure, of personal data [4].

Consent

PIPEDA and the Privacy Act, which govern the private and public (at the federal level) sectors, respectively, differ on the issue of consent: both laws uphold the principle of purpose limitation, requiring, in most circumstances, that an individual's consent is obtained before their personal data can be used for a purpose other than the one for which the data was collected [57, 61]. Only PIPEDA, however, explicitly requires consent to collect personal data in the first place. FR technology used by government agencies exists at the intersection of the public and private sectors: any data framework for FR will likely need different rules regarding consent for data collected by the FR providers vs data collected by the agency itself. Specifically, policy-makers will need to consider the consent requirements for an individual's inclusion in a provider-held training set, a provider-held testing set, an agency or government-held testing set (e.g. for bias verification purposes), a provider-held search dataset (e.g. the database provided to police agencies by Clearview AI), an agency-held search dataset (e.g. a police database), and an agency-held input set (e.g. CCTV footage collected by a police department).

If policy-makers deem that requiring consent in certain contexts and for certain types of data would torpedo the effectiveness and by extension the possible benefits of FR technology (e.g. police departments deploying security cameras on a public street, where meaningful consent would be

impossible to obtain), the public should, at minimum, be informed of its use, in keeping with both the Privacy Act and PIPEDA [57, 62].

The applicability of PIPEDA's consent requirements to provider-held search data under PIPEDA has already been examined by the OPC in the case of Clearview AI's data mining activities. The OPC's conclusion, that Clearview AI did not obtain the requisite consent to amass users' social media photos, has been criticized by some legal experts for its assertions that PIPEDA's exceptions from consent should be interpreted narrowly, particularly its assertion that social media data shouldn't be counted under PIPEDA's "publicly available" exception [91]. It follows that PIPEDA, as well as the Privacy Act, would benefit from updates in order to clarify how they apply to FR.

Bill C-11 aims to update Canada's privacy laws, but legal experts have argued it may do more harm than good, particularly as it relates to consent [4, 68, 85]. The bill largely replicates PIPEDA, including its consent requirements, while empowering the OPC to enforce the new law, primarily through compliance orders to companies [4]. Under PIPEDA, the OPC can only publish non-binding reports [86].

Even under the updated law, however, the OPC's enforcement powers would remain limited. As detailed by Scassa [87], companies would have the ability to appeal not only the OPC's orders but also its findings to a new Personal Information Protection and Data Tribunal, a body that would be required to contain only one expert in privacy law. The OPC would not be able to directly issue fines to offending companies; instead, it would make recommendations to the tribunal. Moreover, only certain breeches could result in fines; a failure to acquire consent, as in the case of Clearview AI, is not one of them, to the consternation of Canada's Privacy Commissioner [5].

Bill C-11 also contains numerous exceptions to both consent *and* knowledge (i.e. the company's duty to inform individuals when it collects or uses their data) that are absent from PIPEDA [85]. For example, "knowledge and consent" are not required in the course of business activities that are "carried out to prevent or reduce the organization's commercial risk" or where "obtaining

the individual's consent would be impracticable because the organization does not have a direct relationship with the individual" [4].

Scassa [87] argued that under the proposed law a company such as Clearview AI would undoubtedly employ the latter exception as a defence. She highlighted, however, the bill's assertion that exceptions to consent are permitted only in situations where "a reasonable person would expect such a collection or use for that activity"; this, she alleged, would not apply to Clearview AI, whose data mining activities have already been deemed inappropriate by the OPC. Nevertheless, despite concluding that the bill as written would likely preclude Clearview AI from prevailing on the basis of such claims, she, like other experts, expressed serious concern over the ambiguity and the large scope of the bill's consent exceptions [4, 87].

Whether or not the new law would be adequate in preventing large scale data collection efforts, it seems nevertheless apparent that such collection efforts are incompatible with most definitions of consent. Compelling an organization like Clearview AI to notify and garner the consent of each individual in their three billion-image database would be infeasible and would likely beget a greater privacy violation than data mining itself. This problem is not limited to particularly egregious cases like Clearview AI's; indeed, a lack of consent is inherent to most modern FR datasets. While Clearview AI's dataset is noteworthy for its unparalleled size, Raji and Fried [76] mapped historical trends in face dataset construction and found that the unprecedented amounts of data required by deep learning has led to the widespread eradication of consent.

Early face datasets were primarily sourced through photo shoots with the full consent of participants. 2007 marked a turning point with the release of the web-scraped Labeled Faces in the Wild dataset and its 13,000 images, which was swiftly followed by numerous other unconstrained (and therefore non-consensual) datasets. The deep learning explosion, which occurred around 2014, triggered an exponential growth in dataset size; images numbered in the millions rather than the thousands, and individual consent became virtually impossible to procure. Raji and Fried also observed that many of these newer datasets contain images of minors or else include descriptive labels

that are racist or sexist. Many of the demographically balanced testing datasets that researchers have touted as being key to solving FR's bias problem, such as Balanced Faces in the Wild (sampled from VGGFace2) and Labeled Ancestral Origin Faces in the Wild (sampled from Bing Image Search API), were assembled without the knowledge or consent of their subjects [2, 80]. It remains unclear how policy-makers should reconcile the principle of consent with the need for large, diverse datasets with which to train and test FR models for accuracy and bias. One approach would be to require consent in order to add a photo to a provider-held search or input dataset, where the risk of material harm is much higher, but not to a provider-held training or testing dataset. Another approach would be to rely on synthetic images, such as the ones used by [42], to construct training and testing datasets.

Right to object and right to be forgotten

Related to the right to consent is the right to object to non-consensual data usage, a principle that is enshrined in Article 21 of the GDPR but that is not currently included in PIPEDA [16, 37]. Both principles are related to the right to be forgotten, i.e. the right to have one's data erased—another right extended to Europeans but not to Canadians [15, 47].

In practice, ensuring that one's likeness is removed from a non-consensual dataset may further violate one's privacy, as demonstrated by Clearview AI's offer to delete an individual's photos only on the condition that they supply the company with a new photo so that their future re-admittance into Clearview AI's database could be prevented [20]. Any organization possessing a continually growing FR dataset for which consent is not prioritized would be unable to permanently exclude an individual from their dataset without the continued possession of some form of unique, personally identifiable data [67]. Storing only the anonymized face embedding associated with the individual's photo would be feasible only if the FR algorithm is immutable.

If, instead, consent was required prior to one's inclusion in an FR dataset, individuals could request that their data be removed without the dataset-holder needing to permanently retain the individual's personal information.

3.3.3 Checklist: what needs to be accomplished before a moratorium can be lifted?

Policy challenges

- Create a data usage framework to establish collection, usage, and retention standards for FR and other biometric data. The framework should account for different types of FR data, should be built on existing data principles such as data minimization, and should define in what circumstances consent is required.

- Update the Privacy Act and PIPEDA and pass new legislation if needed to implement the framework.

- Develop an auditing system to enforce the framework.

3.4 Usage conditions

Even if substantially improved, the Privacy Act and PIPEDA would likely be insufficient in and of themselves in adequately regulating the use of FR by government agencies, as their scope is largely limited to data collection, storage, and use. A new law, similar to the EU's new AI Regulation [14], would likely be needed in order to allow the Canadian government to explicitly identify which use cases of FR should be harshly regulated, and how, and which use cases should be banned outright.

The context in which an FR system is deployed, as well as its specific purpose, have implications far beyond questions of bias or data usage. Different use cases have different risks and benefits, not only for particular groups, but for individuals, FR users, the public, and even for society as a whole. The EU's regulation takes this into account, requiring different levels of auditing and scrutiny for different use cases: developers of high-risk systems will be subjected to periodic auditing to ensure that they comply with various requirements, including the implementation of a risk management system, appropriate data management practices, and record-keeping and technical

documentation, as well as other transparency, oversight, and accuracy conditions [14]. Extremely high-risk systems, such as those that perform live monitoring of public places for the purpose of law enforcement, will be banned except in specific circumstances like terrorist attacks [14].

3.4.1 Key considerations for establishing usage conditions

What constitutes a high-risk use case?

Determining which use cases of FR should be deemed high-risk (or extremely high-risk) is both critical and contentious. The new EU AI Regulation categorizes particular types of AI, such as remote biometric (of which FR is an example) identification systems, as high risk, as they have the potential to violate Europeans' fundamental rights [14]. Other proposed legislation has employed different definitions of "high-risk". The Algorithmic Accountability Act, introduced to the US Congress in 2019, has been criticized for its overly broad definition; for example, under the proposed law, any algorithm that processes gender information would be classified as high-risk [51].

It is also imperative that the list of high-risk and extremely high-risk (i.e. banned) use cases be modifiable over time. The EU AI Regulation allows for new additions by the European Commission to the list of high-risk systems; however, no such rule is explicitly laid out for banned use cases [14].

Social validity and society-wide risks

How do we measure the social validity and society-wide risk of an FR use case? We have so far examined harm stemming from the FR system itself (e.g. bias originating from the model or from the training data), as well as harm stemming from how the model is used and the environment in which it is used (e.g. police departments exacerbating bias, data collectors overlooking consent). In determining for which use cases of FR the potential benefits outweigh the risk, we must situate the FR system in its larger social context and consider its long-term effects on Canadian society, not only on individuals directly impacted by its decisions. A moratorium would allow researchers

and policy-makers to investigate the ultimate question: even if the problem of bias is solved, are there uses of FR that carry too much societal risk to be engaged in by government agencies?

As explained above, meaningful consent is incompatible with many aspects of FR, such as its use in public places, as well as large-scale data collection for training, testing, and bias mitigation. The former is of heightened concern, as such practices amount to surveillance that is more pervasive and more acute than any to which most Canadians have previously been subjected. While Canadians are currently protected by strong democratic institutions, norms, and laws, albeit ones that are somewhat lax when applied to FR, there is no definitive guarantee that Canadian democracy will remain stable decades into the future. Indeed, the previous decade and a half has been characterized by a global decline in the number of democracies and in democratic freedoms [78]. If data minimization principles are not followed, future governments will be able to easily obtain access to decades-worth of biometric data. Even if such principles are upheld, if FR is widely and freely used by the police and other government agencies, the tools and practices necessary to implement an autocratic surveillance state will be readily available.

While this scenario may appear inconceivable, there is evidence that surveillance has severely negative effects, even when integrated into an otherwise democratic society. Researchers have argued that heavy surveillance results in self-censoring and dissuades people from exercising their rights [90]. One example is the progression of LGBTQ rights; within 40 years, Canada transformed from a society in which sex between men was a prosecutable offence to one which largely embraced same sex marriage. Experts argue that privacy and anonymity, and the countercultures they protect, are necessary for such changes to occur [90].

Thus, before a moratorium can be lifted, researchers and policy-makers should aim to better understand the effects of FR surveillance on the Canadian public in order to assess whether certain use cases of the technology (e.g. widespread public use of FR by government agencies) pose too large of a threat to be authorized.

Existing principles beyond risk mitigation

While assessing potential risks is critical when deciding whether to ban or regulate certain use cases of FR, other ethical principles, such as proportionality, should likewise be taken into consideration.

A key example is the case of the Swedish school that was fined for implementing an FR-based attendance system; the Swedish data protection authority's decision was based in part on how disproportionate the solution (applying FR) was to the problem (tracking attendance) [24].

Under the proposed EU AI Regulation, the principle of proportionality is utilized when assessing whether certain use cases, namely exceptions to the ban on real-time biometric identification, are appropriate [14]. The seriousness of the crime, the likelihood and degree of potential harm from FR, and the context in which the FR system would be used are all considered when determining if a particular exception is proportionate.

Notably, the regulation does not extend this proportionality condition to high-risk or low-risk systems [14]. In other words, while the proposed regulation uses proportionality to justify exceptions to its ban on the most dangerous type of FR, the reverse is not true: authorized use cases that violate the principle of proportionality do not appear to be given any special consideration. Many uses of FR that do not present a high enough risk to be banned under the proposed regulation might nevertheless violate the principle of proportionality. For example, relying on extensive facial recognition searches to catch the perpetrators of low-level crimes like shoplifting is arguably a disproportionate response. Canadian policy-makers would be advised to examine the proportionality of each use case, in addition to its risk level, when deciding whether to ban, regulate, or authorize.

Restricting access and frequency

In addition to banning extremely high-risk uses of FR, policy-makers could also curb the ease with which public agencies engage in high-, medium-, or even, when appropriate, low-risk uses of the technology. By introducing authorization requirements, the frequency and overall risk of FR usage could be greatly dampened.

In December 2020, motivated in large part by an ACLU investigation that uncovered widespread and unrestricted FR usage by Massachusetts police, state senators attempted to pass strict FR reforms but were blocked by the state's Republican governor [34]. A compromise was reached whereby a judge's permission must be acquired before local police can request that an FR search of the state police, FBI, or Registry of Motor Vehicles database be performed.

3.4.2 Checklist: what needs to be accomplished before a moratorium can be lifted?

Policy challenges

- Implement a law similar to the EU's new AI Regulation [14] that will allow the government to ban, regulate, and authorize particular use cases of FR, in a way that accounts for the risk level of different use cases, as well as other principles such as proportionality.

Research challenge

- Research the societal impacts of particularly high-risk use cases of FR such as widespread public surveillance.

Chapter 4

Bias mitigation through calibration

The method and experiments presented in this chapter are taken from a paper by Salvador et al. [83] that is currently under submission. As the paper's second author, I proposed the central problem of fairness in facial recognition and pointed out that calibration had not been previously explored for FR. I conducted a thorough literature review (as shown in section 2.2, subsection 3.2.1, and subsection 3.2.6) and through many discussions helped design experiments to show the effectiveness of our method. I also contributed to the writing and editing of the paper. The experiments were coded by Tiago Salvador with help from Noah Marshall. Tiago Salvador ran the experiments and prepared the figures and tables contained in this chapter. For a full discussion of methods and results, see [83].

4.1 Motivation

Solving facial recognition's bias problem is arguably the most important condition for ensuring its safe use. As we saw in Chapter 3, researchers have applied a myriad of fairness definitions when assessing FR bias. However, to the best of our knowledge, no work thus far has applied the metric of fairness-calibration to FR. As Pleiss et al. [72] explained, fairness-calibration is a crucial condi-

tion to ensure fairness in many contexts. For example, if a model to predict a defendant's recidivism risk is not fairly-calibrated between racial groups, a Black defendant and a white defendant with the same predicted score may have different true probabilities of reoffending, thus inciting judges to account for race in their sentencing decisions. The same danger of making race-based decisions extends to uncalibrated facial recognition, particularly when it is utilized for policing or sentencing. To rectify this, Salvador et al. [83] introduced a simple way to ensure fairness-calibration while likewise maintaining equally low FPRs between demographic groups. Recall that perfectly achieving the three conditions of fairness-calibration, predictive equality (i.e. equal FPRs), and equal opportunity (i.e. equal FNRs) is impossible except in very narrow circumstances [12, 41].

Our bias mitigation approach has three considerable advantages. Firstly, unlike many methods described in Chapter 3, it is not computationally expensive, nor does it require re-training an FR model; rather, it can be easily appended to an existing system. Secondly, our method does not depend on knowledge of demographic groups; variations of the approach can be applied whether or not demographic labels are present in the search dataset. Finally, it achieves both fairness-calibration and predictive equality (beating comparable post-hoc methods) without sacrificing accuracy; in fact, it increases a model's accuracy.

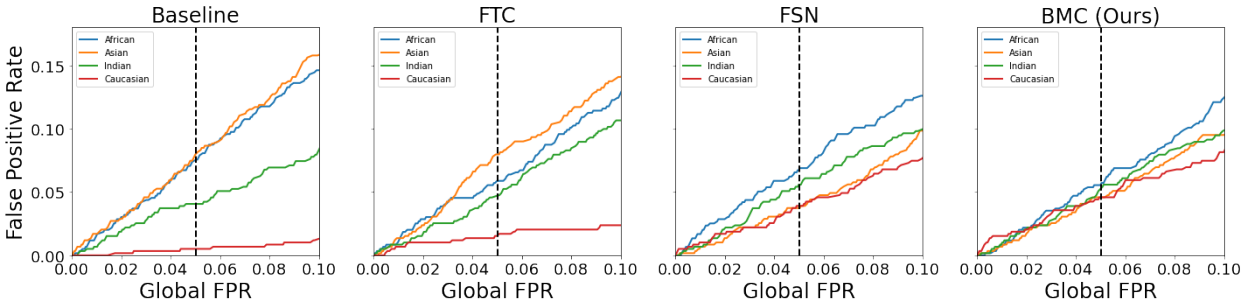

Figure 4.1: Reduction in bias on RFW using the FaceNet (WebFace) model, as measured by group FPRs. Our Bias Mitigation Calibration (BMC) method results in lower bias (lines closer together) than comparable post-hoc methods FTC [96] and FSN [95]. Figure source: [83]

Our method also allows users of the FR system to choose, based on the use context, which fairness condition (predictive equality or equal opportunity) they wish to fulfill in addition to fairness-

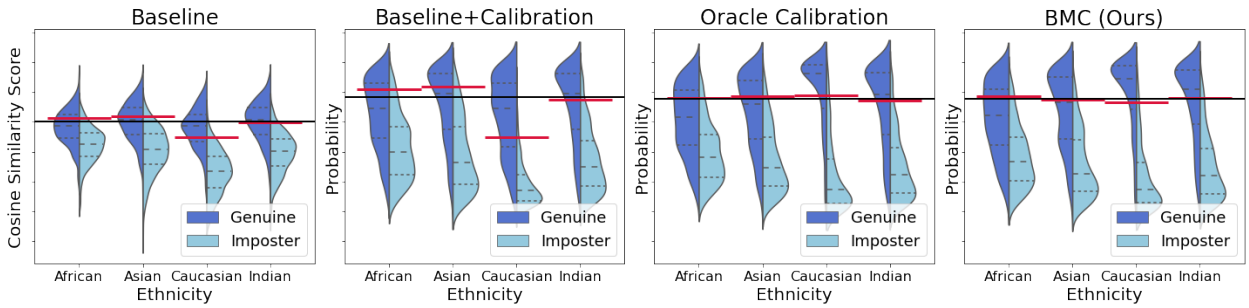

Figure 4.2: Bias reduction, as measured by deviation between the black line (threshold needed to achieve a global FPR of 5%) and red lines (thresholds needed to achieve FPRs of 5% for each demographic group). Simply applying a standard calibration method does not reduce bias. Our methods, Oracle Calibration (separately calibrating each demographic group; group membership must be known) and BMC (separately calibrating clusters that were generated using feature vectors), successfully reduce bias. Figure source: [83]

calibration, simply by setting a low global FPR (if predictive equality is desired) or a low global FNR (if equal opportunity is desired). For instance, equal FPRs should be prioritized when the potential for harm stemming from a false positive (e.g. a false arrest or conviction) is substantial. Conversely, for an office building equipped with an FR security system, a slew of false negatives would prove highly disruptive for workers; equal FNRs should thus be given priority.

4.2 Calibration vs fairness-calibration

Recall that fairness-calibration is satisfied if for any outputted score $s \epsilon [0, 1]$:

$$P(Y = 1|S = s, G = a) = P(Y = 1|S = s, G = b) = s \tag{4.1}$$

That is, fairness-calibration holds if, regardless of group membership, any image pair with a score s truly has a probability s of being a genuine match.

Calibration, for its part, holds if a pair with a score s has an *overall* probability s of being a genuine match, i.e. for any $s \epsilon [0, 1]$:

$$P(Y = 1|S = s) = s \tag{4.2}$$

A model may be calibrated but not fairly-calibrated. For example, consider two groups of equal size, both of whose members are all given scores of 0.5. If group A's members all have true probabilities of 0.4, and group B's members all have true probabilities of 0.6, then the model satisfies the notion of calibration but does not satisfy fairness-calibration.

4.3 Standard calibration methods

We can apply a standard approach to ensure overall calibration (i.e. to ensure that predicted scores correspond to real probabilities); however, as we just saw, this is insufficient to guarantee that an FR model is fairly-calibrated between different demographic groups. An additional step is required, wherein we divide the dataset into groups or clusters and separately apply a calibration method to each one.

We begin by presenting traditional calibration methods, any of which could be adopted for our fairness-calibration approach. Each of the below methods converts pairs' similarity scores into confidence scores that better reflect the pair's true probability of depicting the same person.

First, we define $P^{\text{cal}} \subseteq X \times X$ as the set of image pairs that we aim to calibrate. We designate S^{cal} to be the set of corresponding cosine similarity scores. That is, for a pair $(x_i, x_j) \in P^{\text{cal}}$, we set $s(x_i, x_j) = \frac{f(x_1) \cdot f(x_2)}{\|f(x_1)\| \|f(x_2)\|} \in S^{\text{cal}}$, where $f(x_i)$ is the face embedding generated by the FR model. We likewise define $I(x_i)$ as the identity of image i and $Y = y_{i,j}$ as the true class (1 or 0, i.e. genuine pair or not) of a pair (x_i, x_j).

4.3.1 Histogram binning

Histogram binning [108] is a simple method that involves dividing S^{cal} into N bins. This division can be made to ensure that the bins have equal weight or are equally spaced, or even to maximize mutual information [69]. Each similarity score $s(x_i, x_j) \in S^{\text{cal}}$ is assigned to one bin B_n, where $n = 1, ...N$. The score is then converted into a confidence score that reflects the proportion of scores in B_n that truly correspond to genuine pairs. In other words, for a pair (x_i, x_j) whose score

$s(x_i, x_j)$ belongs to B_n, we assign the pair the following confidence score:

$$c_{ij} = \frac{1}{|B_n|} \sum_{\substack{s_{lm} \in B_i \\ (x_l, x_m) \in P^{\text{cal}}}} \mathbb{1}_{I(x_l)=I(x_m)} \tag{4.3}$$

4.3.2 Isotonic regression

Isotonic regression [54] involves learning a confidence function $c : \mathbb{R} \to \mathbb{R}$ by solving

$$\underset{\mu}{\text{argmin}} \frac{1}{|P^{\text{cal}}|} \sum_{(x_l, x_m) \in P^{\text{cal}}} \left(\mu(s(x_l, x_m)) - \mathbb{1}_{I(x_l)=I(x_m)}\right)^2 \tag{4.4}$$

and then setting $c(x_i, x_j) = \mu(s(x_i, x_j))$.

4.3.3 Beta calibration

Beta calibration [69], a generalization of logistic regression, involves learning the calibration map

$$c_\theta(s) = \mu(s; \theta_1, \theta_2, \theta_3) = \frac{1}{1 + 1/\left(e^{\theta_3} \frac{s^{\theta_1}}{(1-s)^{\theta_2}}\right)} \tag{4.5}$$

where $\theta_1, \theta_2, \theta_3$ are parameters selected by minimizing

$$LL(c, y) = y(-\log(c)) + (1 - y)(-\log(1 - c)) \tag{4.6}$$

where $c = \mu(s(x_i, x_j))$.

4.4 Applying fairness-calibration methods to facial recognition

4.4.1 Oracle Calibration

If demographic labels are present, the search dataset can readily be divided into groups based on a function A that assigns an attribute a (e.g. gender) to each image. We can then convert

the similarity scores for each group into confidence scores by applying a standard calibration approach (like the ones discussed in section 4.3) to each individual group. So, given a pair of images $(x_i, x_j) \in P^{\text{cal}}$ with a score $s(x_i, x_j)$, we set the pair's confidence score as:

$$c(x_i, x_j) = \begin{cases} c_a(s(x_i, x_j)) & \text{if } A(x_i) = A(x_j) = a \\ 0 & \text{if } A(x_i) \neq A(x_j) \end{cases} \tag{4.7}$$

where c_a is the confidence function for group a. Salvador et al. [83] dubbed this method 'Oracle calibration', as it requires knowledge of demographic information. In practice, the usability of this method would be limited, particularly when mitigating racial bias, as racial identity cannot be precisely partitioned into discrete categories.

4.4.2 Bias Mitigation Calibration

In the absence of demographic labels, we can instead employ a k-means algorithm (with $k = 100$, as determined through testing) to generate k distinct clusters of images. As with the Oracle approach, calibration can be achieved for each group (cluster) via a standard calibration method. However, unlike with Oracle calibration, it is possible that two images from different clusters do indeed form a genuine pair. Thus, for a pair $(x_i, x_j) \in P^{\text{cal}}$ in clusters a and b, respectively, we set its confidence score as:

$$c(x_i, x_j) = \theta c_a(s(x_i, x_j)) + (1 - \theta) c_b(s(x_i, x_j)),$$

where c_a is the confidence function for cluster a and θ is the population fraction of cluster a relative to cluster b.

4.5 Experiments

4.5.1 Models and datasets

We tested our method by appending it to three existing FR models: an ArcFace model [89] trained on a refined version of the MS-Celeb-1M dataset [30], and two Inception ResNet models [25] trained on the VGGFace2 [10] and CASIA Web-face [106] datasets. We employed two demographically-labeled testing datasets: Racial Faces in the Wild (RFW) [102] and Balanced Faces in the Wild (BFW) [80]. Prior to testing, we pre-processed our images using a Multi-Task Convolutional Neural Network (MTCNN).

4.5.2 Methods compared

We compared our Oracle and Bias Mitigation Calibration (using k = 100) methods to a baseline method that applied beta calibration [69] to the dataset as a whole (instead of to individual groups). We also tested two post-hoc bias mitigation methods introduced by other researchers, namely Terhörst et al.'s fair template comparison (FTC) method [96] and Terhörst et al.'s fair score normalization (FSN) method [95]. To make a fair comparison, we applied (overall) beta calibration to both of these methods. Table 4.1 shows the advantages of our method over FTC and FSN.

Table 4.1: Advantages and disadvantages of the different post-hoc bias mitigation methods. Table source: [83]

Method	Improves accuracy over Baseline	Predictive equality (equal FPRs)	Fairly-calibrated	Does not require sensitive attribute during training	Does not require sensitive attribute at test time
FTC [96]	✔	✔	✔	✔	✔
FSN [95]	✔	✔	✔	✔	✔
Oracle (Ours)	✔	✔	✔	✔	✔
BMC (ours)	✔	✔	✔	✔	✔

4.5.3 Metrics

To test for fairness-calibration, we measured the calibration error on each demographic group using a metric inspired by the Kolmogorov-Smirnov (KS) statistic test and introduced by Gupta et al. [31]. Their metric compares the cumulative distributions of $P(Y = 1, C = c)$ and $cP(C = c)$,

which correlate with calculating the sequences $\{h_n\}$ and $\{\tilde{h}_n\}$. Let N be the total number of pairs such that for all $n \in N$, (x_{n_i}, x_{n_j}) is a pair of images. Then

$$h_n = h_{n-1} + \mathbb{1}_{y_{n_i,n_j}=1}/N \quad \text{and} \quad \tilde{h}_n = \tilde{h}_{n-1} + c(x_{n_i}, x_{n_j})/N \tag{4.8}$$

where $h_0 = \tilde{h}_0 = 0$. The KS calibration error metric is then set as

$$KS = \max_n |h_n - \tilde{h}_n|. \tag{4.9}$$

4.5.4 Results

We found that our Bias Mitigation Calibration (BMC) method was the best, of the post-hoc methods surveyed, at global accuracy, fairness-calibration, and predictive equality. We used 5-fold cross validation to generate our results. See the appendix of [83] for more detailed results, including the standard deviation errors for each method.

Global accuracy

We measured global accuracy by computing the area under the ROC curve[1] and the TPR at different global FPRs (0.1% and 1%) (see Table 4.2). Our BMC method obtained the best AUROC in all cases and the highest TPR in seven out of eight cases.

Fairness-calibration

We wanted to ensure that a pair's outputted probability matched its true probability of being a match, regardless of which demographic group(s) it belonged to. To this end, we measured each group's calibration error (KS). Moreover, we wanted to minimize the deviations between groups' calibration errors, as measured by the Average Absolute Deviation (AAD), the Maximum Absolute Deviation (MAD), and the Standard Deviation (SD); see Table 4.3.

[1]Note that while subsection 3.2.1 described the limitations of the AUROC, this was in reference to its use in measuring bias, not in measuring overall accuracy, which is indeed its original function.

BMC attained the lowest mean calibration error (across groups) in three of the four cases and the lowest deviation in all cases (see Table 4.3).

Predictive equality

We tested how close the different methods were to reaching perfect predictive equality by computing the deviation (AAD, MAD, and STD) between demographic groups' FPRs at global FPRs of 0.1% and 1%.

Our BMC method came out on top in 18 of 24 cases (see Table 4.4). In the remaining six cases, FTC performed best. However, unlike BMC, FTC's reduction in bias was achieved only at the expense of a decrease in overall accuracy (see Table 4.2).

Context-dependent bias mitigation: Equal opportunity

Our method can be readily modified (simply by setting a low global FNR instead of a low global FPR) so as to be applied to contexts where equal opportunity is more valuable than predictive equality. Table 4.5 shows that in most cases BMC achieved comparable results to its rival methods.

Tables

Table 4.2 shows the global accuracy of the different methods across the different models, while Tables 4.3-4.5 show how well the various methods satisfied the conditions of fairness-calibration, predictive equality, and equal opportunity. The best results for each FR model are bolded. Since our Oracle method requires knowledge of the sensitive attribute, it is largely infeasible in practice; we therefore did not consider it when determining the best method. Note, however, that our BMC method beat or was comparable to our Oracle method by nearly all measures.

Table 4.2: Global accuracy, as measured by the AUROC and the TPR at global FPRs of 0.1% and 1%. Table source: [83]

	RFW						BFW					
	FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
(↑)	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR	AUROC	TPR @ 0.1% FPR	TPR @ 1% FPR
Baseline	88.26	18.42	34.88	83.95	11.18	26.04	96.06	33.61	58.87	97.41	86.27	90.11
FTC	86.46	6.86	23.66	81.61	4.65	18.40	93.30	13.60	43.09	96.41	82.09	88.24
FSN [95]	90.05	23.01	40.21	85.84	17.33	32.80	96.77	**47.11**	68.92	97.35	86.19	90.06
BMC (Ours)	**90.58**	**23.55**	**41.88**	**86.71**	**20.64**	**33.13**	**96.9**	46.74	**69.21**	**97.44**	**86.28**	**90.14**
Oracle (Ours)	*89.74*	*21.4*	*41.83*	*85.23*	*16.71*	*31.6*	*97.28*	*45.13*	*67.56*	*98.91*	*86.41*	90.40

Table 4.3: Fairness-calibration is achieved if two conditions are met: 1) all subgroups have a low calibration error (as measured by the mean KS across demographic subgroups), and 2) the calibration error differs little between subgroups (as measured by the Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD) of KS across subgroups). Table source: [83]

	RFW								BFW							
	FaceNet (VGGFace2)				FaceNet (Webface)				FaceNet (Webface)				ArcFace			
(↓)	Mean KS	AAD	MAD	STD	Mean KS	AAD	MAD	STD	Mean KS	AAD	MAD	STD	Mean KS	AAD	MAD	STD
Baseline	6.37	2.89	5.73	3.77	5.55	2.48	4.97	2.91	6.77	3.63	5.96	4.03	2.57	1.39	2.94	1.63
FTC [96]	5.16	2.31	4.44	2.87	4.11	1.87	3.74	2.20	6.60	2.42	5.19	2.93	3.70	1.50	3.07	1.78
FSN [95]	1.43	0.35	0.57	0.40	2.49	0.84	1.19	0.91	**2.76**	1.38	2.67	1.60	2.65	1.45	3.23	1.71
BMC (Ours)	**1.37**	**0.28**	**0.5**	**0.34**	**1.75**	**0.41**	**0.64**	**0.45**	3.09	**1.34**	**2.48**	**1.55**	**2.49**	**1.3**	**2.68**	**1.52**
Oracle (Ours)	*1.18*	*0.28*	*0.53*	*0.33*	*1.35*	*0.38*	*0.66*	*0.43*	*2.23*	*1.15*	*2.63*	*1.4*	*1.41*	*0.59*	*1.3*	0.69

Table 4.4: Predictive equality is achieved if the demographic groups have equal FPRs. We thus measure the Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD) between the groups' FPRs at fixed global FPRs of 0.1% and 1%. Table source: [83]

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
		AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FPR	Baseline	0.10	0.15	**0.10**	0.14	0.26	0.16	0.29	1.00	0.40	0.12	0.30	0.15
	FTC FTC	0.10	0.15	0.11	0.12	0.23	0.14	0.24	0.74	0.32	**0.09**	**0.20**	**0.11**
	FSN [95]	0.10	0.18	0.11	0.11	0.23	0.13	**0.09**	**0.20**	**0.11**	0.11	0.28	0.14
	BMC (Ours)	**0.09**	**0.14**	**0.10**	**0.09**	**0.16**	**0.1**	**0.09**	**0.20**	**0.11**	0.11	0.31	0.15
	Oracle (Ours)	*0.11*	*0.19*	*0.12*	*0.11*	*0.2*	*0.13*	*0.12*	*0.25*	*0.15*	*0.12*	*0.27*	*0.14*
1% FPR	Baseline	0.68	1.02	0.74	0.67	1.23	0.79	2.42	7.48	3.22	0.72	1.51	0.85
	FTC [96]	0.60	0.91	0.66	0.54	1.05	0.66	1.94	5.74	2.57	**0.54**	**1.04**	**0.61**
	FSN [95]	0.37	0.68	0.46	0.35	0.61	0.40	0.87	2.19	1.05	0.55	1.27	0.68
	BMC (Ours)	**0.28**	**0.46**	**0.32**	**0.29**	**0.57**	**0.35**	**0.8**	**1.79**	**0.95**	0.63	1.46	0.78
	Oracle (Ours)	*0.4*	*0.69*	*0.45*	*0.41*	*0.74*	*0.48*	*0.77*	*1.71*	*0.91*	*0.83*	*2.08*	*1.07*

Table 4.5: Equal opportunity is achieved if the demographic groups have equal FNRs. We thus measure the Average Absolute Deviation (AAD), Maximum Absolute Deviation (MAD), and Standard Deviation (STD) between the groups' FNRs at fixed global FNRs of 0.1% and 1%. Table source: [83]

		RFW						BFW					
		FaceNet (VGGFace2)			FaceNet (Webface)			FaceNet (Webface)			ArcFace		
	(↓)	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD	AAD	MAD	STD
0.1% FNR	Baseline	0.09	0.13	0.10	0.10	0.16	0.11	0.09	0.23	0.11	0.11	0.31	0.14
	FTC [96]	0.09	**0.11**	**0.09**	**0.08**	0.14	**0.1**	**0.04**	**0.09**	**0.05**	**0.06**	**0.14**	**0.07**
	FSN [95]	**0.09**	0.13	0.09	0.09	**0.14**	0.10	0.07	0.22	0.10	0.12	0.33	0.15
	BMC (Ours)	0.10	0.14	0.10	0.11	0.17	0.12	0.10	0.27	0.13	0.09	0.17	0.10
	Oracle (Ours)	*0.11*	*0.18*	*0.12*	*0.12*	*0.21*	*0.13*	*0.09*	*0.24*	*0.11*	*0.11*	*0.32*	0.14
1% FNR	Baseline	0.60	0.96	0.67	0.45	0.81	0.53	0.39	0.84	0.47	0.75	1.85	0.93
	FTC [96]	0.48	0.83	0.56	**0.32**	**0.58**	**0.38**	**0.3**	**0.62**	**0.34**	**0.49**	**1.12**	**0.6**
	FSN [95]	**0.28**	**0.47**	**0.32**	0.40	0.78	0.48	0.41	0.92	0.49	0.77	1.91	0.96
	BMC (Ours)	0.30	0.51	0.34	0.39	0.72	0.48	0.32	0.74	0.40	0.65	1.48	0.80
	Oracle (Ours)	*0.38*	*0.61*	*0.42*	*0.56*	*1.06*	*0.67*	*0.37*	*0.77*	*0.44*	*0.5*	*1.11*	0.60

Chapter 5

Conclusion

5.1 Summary

The aim of this thesis was twofold: firstly, to examine the conditions, both technical and policy-oriented, required to lift a potential moratorium on the use of facial recognition technology by federal government agencies, and secondly, to present a promising step toward fulfilling one of the most fundamental of these conditions: context-dependent bias mitigation.

In Chapter 2, we described how facial recognition technology works, how it is used in Canada, and why, due to concerns around bias and privacy, it poses a risk to Canadians. We then presented a selection of key fairness definitions used to assess bias present in FR models. We recreated a proof from [12] demonstrating that of the following three fairness definitions—predictive equality (i.e. equal false positive rates), equal opportunity (i.e. equal false negative rates), and fairness-calibration—no more than two can be satisfied simultaneously, except under strict conditions.

Next, in Chapter 3, we explored the conditions that should be met before policy-makers consider removing a moratorium on government use of FR. We grouped these conditions into three categories: accuracy and bias conditions, data conditions, and usage conditions.

In our section on accuracy and bias, we highlighted one primary policy condition: the development of an auditing system to assess FR bias. We explained that this auditing system should inspect not only the model itself, but how it was built and how it is used (e.g. which fairness definition(s) is most relevant, which demographic groups may be most harmed by the model's use, etc.). We also detailed the following research challenges: 1) the creation of context-based bias mitigation solutions; 2) an analysis of the relevance of different fairness definitions for different use cases; 3) a study of FR bias against overlooked demographic groups; 4) an investigation into how to evaluate bias in cases when discrete and mutually exclusive groups are nonexistent; and finally, 5) an examination of the technical and legal implications of prioritizing some groups over others when assessing bias.

We laid out one primary data-related condition: the creation of a data usage framework—one that is based on key principles such as data minimization and that clearly addresses consent—to regulate the use, collection, and retention of FR data. As we explained, implementing such a framework would necessitate modifying (or replacing) laws like the Privacy Act and PIPEDA. An auditing system to assess compliance would likewise be needed.

We also outlined the following usage condition: the implementation of a law to outlaw or regulate different uses of FR, depending on use cases' level of risk and their adherence to principles such as proportionality. We also stressed the need for researchers to scrutinize and deepen our understanding of the society-wide risks of certain FR use cases such as mass surveillance.

Finally, in Chapter 4, we tackled the challenge of bias mitigation, introducing our Bias Mitigation Calibration method [83], a post-hoc method that allows FR models to successfully satisfy both fairness-calibration and predictive equality, while simultaneously increasing their overall accuracy. Our method has numerous advantages: 1) it can be applied to any existing FR model with no retraining required; 2) it does not require that images possess demographic labels; 3) it achieves better accuracy and lower bias than comparable post-hoc methods; and 4) if called for by the use

case and deployment context, it could instead be used to achieve fairness-calibration and equal opportunity (as opposed to fairness-calibration and predictive equality).

5.2 Key takeaways and future work

Facial recognition technology, as it currently exists, is biased and dangerous. The threat it poses both to individuals and to our democratic society cannot be mitigated by technical approaches or by policy approaches alone. A moratorium on government usage of FR would give researchers time to tackle FR systems' technical failings and to study the true impacts of their use. Our work demonstrates that with a solid understanding of fairness metrics, models can indeed be made substantially fairer. A moratorium would also allow policy-makers to implement sound, technically informed regulations—and if needed, outright bans. Addressing the problem of bias, for instance, will require designing auditing systems that account for different context-dependent fairness definitions.

Future work should consider the practical considerations of implementing a moratorium on the use of FR by government agencies, as well as the feasibility of extending a similar moratorium to the private sector and to public agencies at the provincial level. A future project could also build on the work in this thesis, as well as on [67], to develop a comprehensive research agenda to be undertaken during a moratorium. Such a project could likewise examine the role public consultation, particularly with marginalized groups, should play in crafting appropriate policy solutions.

Bibliography

[1] Kate Allen, Wendy Gillis, and Alex Boutillier. Facial recognition app Clearview AI has been used far more widely in Canada than previously known. https://www.thestar.com/news/canada/2020/02/27/facial-recognition-app-clearview-ai-has-been-used-far-more-widely-in-canada-than-previously-known.html, February 2020.

[2] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellaker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision (EECV)*, pages 556–572, 2019.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Technical report, Propublica, May 2016.

[4] Lisa Austin. Who decides? Consent, meaningful choices, and accountability. https://srinstitute.utoronto.ca/news/austin-consent-meaningful-choice-accountability, December 2020.

[5] Ian Bailey. Privacy Commissioner calls for bill to include tougher regulation of facial recognition technology. https://www.theglobeandmail.com/politics/article

-privacy-commissioner-calls-for-bill-to-include-tougher-regulation-of/, May 2021.

[6] Axon AI & Policing Technology Ethics Board. First report of the Axon AI & Policing Technology Ethics Board. https://static1.squarespace.com/static/58a33e881b631bc60d4f8b31/t/5d13d7e1990c4f00014c0aeb/1561581540954/Axon_Ethics_Board_First_Report.pdf, June 2019.

[7] Owen Bowcott. UK's facial recognition technology 'breaches privacy rights'. https://www.theguardian.com/technology/2020/jun/23/uks-facial-recognition-technology-breaches-privacy-rights, June 2020.

[8] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018.

[9] Sheri Byrne-Haber. Disability and AI bias. https://sheribyrnehaber.medium.com/disability-and-ai-bias-cced271bd533, July 2019.

[10] Q. Cao, Li Shen, Weidi Xie, O. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74, 2018.

[11] Jacqueline G Cavazos, P. Jonathon Phillips, Carlos D. Castillo, and Alice J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Trans Biom Behav Identity Sci.*, 3(1):101–111, January 2021.

[12] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, June 2017.

[13] Victoria Clayton. The problem with the GRE. https://www.theatlantic.com/education/archive/2016/03/the-problem-with-the-gre/471633/, March 2016.

[14] European Commission. Proposal for a regulation laying down harmonised rules on artificial intelligence. Technical report, European Commission, April 2021.

[15] Intersoft Consulting. Art. 17 GDPR: Right to erasure ('right to be forgotten'). https://gdpr-info.eu/art-17-gdpr/.

[16] Intersoft Consulting. Art. 21 GDPR: Right to object. https://gdpr-info.eu/art-21-gdpr/.

[17] Intersoft Consulting. Art. 5 GDPR: Principles relating to processing of personal data. https://gdpr-info.eu/art-5-gdpr/.

[18] Intersoft Consulting. Art. 9 GDPR: Processing of special categories of personal data. https://gdpr-info.eu/art-9-gdpr/.

[19] Cynthia M. Cook, John J. Howard, Yevgeniy B. Sirotin, Jerry L. Tipton, and Arun R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, January 2019.

[20] Thomas Daigle. Canadians can now opt out of Clearview AI facial recognition, with a catch. https://www.cbc.ca/news/technology/clearview-ai-canadians-can-opt-out-1.5645089, July 2020.

[21] Thomas Daigle. Clearview AI facial recognition offers to delete some faces — but not in Canada. https://www.cbc.ca/news/technology/clearview-ai-canadian-data-1.5605258, June 2020.

[22] Prithviraj Dhar, Joshua Gleason, Hossein Souri, Carlos D. Castillo, and Rama Chellappa. Towards gender-neutral face descriptors for mitigating bias in face recognition. https://arxiv.org/abs/2006.07845, 2020.

[23] Pam Dixon. A failure to do no harm – India's Aadhaar biometric ID program and its inability to protect privacy in relation to measures in Europe and the U.S. *Health and Technology*, 7(4):539–567, June 2017.

[24] Sofia Edvardsen. How to interpret Sweden's first GDPR fine on facial recognition in school. https://iapp.org/news/a/how-to-interpret-swedens-first-gdpr-fine-on-facial-recognition-in-school/, August 2019.

[25] Tim Esler. Face recognition using Pytorch. https://github.com/timesler/facenet-pytorch, 2021.

[26] Robson Fletcher. Calgary police now admit 2 officers used controversial Clearview AI facial-recognition software. https://www.cbc.ca/news/canada/calgary/calgary-police-admit-using-clearview-ai-facial-recognition-software-1.5480803, February 2020.

[27] Chris Frey. Revealed: how facial recognition has invaded shops – and your privacy. https://www.theguardian.com/cities/2016/mar/03/revealed-facial-recognition-software-infiltrating-cities-saks-toronto, March 2016.

[28] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Jointly de-biasing face recognition and demographic attribute estimation. In *European Conference on Computer Vision (ECCV)*, pages 330–347, 2020.

[29] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test part 3: Demographic effects. Technical report, National Institute of Standards and Technology, December 2019.

[30] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (EECV)*, 2016.

[31] Kartik Gupta, Amir Rahimi, Thalaiyasingam Ajanthan, Thomas Mensink, Cristian Sminchisescu, and Richard Hartley. Calibration of neural networks using splines. In *International Conference on Learning Representations*, 2021.

[32] Rebecca Heilweil. The world's scariest facial recognition company, explained. https://www.vox.com/recode/2020/2/11/21131991/clearview-ai-facial-recognition-database-law-enforcement, May 2020.

[33] Kashmir Hill. Another arrest, and jail time, due to a bad facial recognition match. https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail.html, December 2020.

[34] Kashmir Hill. How one state managed to actually write rules on facial recognition. https://www.nytimes.com/2021/02/27/technology/Massachusetts-facial-recognition-rules.html, February 2021.

[35] Kashmir Hill and Gabriel J.X. Dance. Clearview's facial recognition app is identifying child victims of abuse. https://www.nytimes.com/2020/02/07/business/clearview-facial-recognition-child-sexual-abuse.html, February 2020.

[36] Daniel E. Ho, Emily Black, Maneesh Agrawala, and Fei-Fei Li. Domain shift and emerging questions in facial recognition technology. Technical report, Stanford University Human-Centered Artificial Intelligence, November 2020.

[37] Amba Kak and Rashida Richardson. The Office of the Privacy Commissioner of Canada consultation: Proposals for ensuring appropriate regulation of artificial intelligence. Technical report, AI Now, March 2020.

[38] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 2564–2572, July 2018.

[39] Os Keyes. The misgendering machines: Trans/HCI implications of automatic gender recognition. *Proceedings of the ACM on Human-Computer Interaction*, November 2018.

[40] Os Keyes. Gender classification and bias mitigation: a post-publication review. `https://ironholds.org/debiasing/`, July 2020.

[41] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *The 8th Innovations in Theoretical Computer Science Conference*, 2017.

[42] Adam Kortylewski, Bernhard Egger, Andreas Schneider, Thomas Gerig, Andreas Morel-Forster, and Thomas Vetter. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2261–2268, 2019.

[43] KS Krishnapriya, Kushal Vangara, Michael C. King, Vitor Albiero, and Kevin Bowyer. Characterizing the variability in face recognition accuracy relative to race. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, pages 2278–2285, 2019.

[44] Vijay Kumar, R. Raghavendra, Anoop M. Namboodiri, and Christoph Busch. Robust transgender face recognition: Approach based on appearance and therapy factors. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, page 1–7, February 2016.

[45] Sébastien Louradour. What to know about the EU's facial recognition regulation – and how to comply. `https://www.weforum.org/agenda/2021/04/facial-recogn`

ition-regulation-eu-european-union-ec-ai-artificial-intelligence-machine-learning-risk-management-compliance-technology-providers/, April 2021.

[46] A. Lumaka, N. Cosemans, A. Lulebo Mampasi, G. Mubungu, N. Mvuama, T. Lubala, S. Mbuyi-Musanzayi, J. Breckpot, M. Holvoet, T. de Ravel, G. Van Buggenhout, H. Peeters, D. Donnai, L. Mutesa, A. Verloes, P. Lukusa Tshilobo, and K. Devriendt. Facial dysmorphism is influenced by ethnic background of the patient and of the evaluator. *Clinical Genetics*, 92(2):166–171, February 2017.

[47] Mairead Matthews. Facial recognition company Clearview AI provides a useful case study for the right to be forgotten in Canada. https://medium.com/digitalthinktankictc/facial-recognition-company-clearview-ai-provides-a-useful-case-study-for-the-right-to-be-forgotten-1b2584065e2e, June 2020.

[48] Michael McLaughlin and Daniel Castro. The critics were wrong: NIST data shows the best facial recognition algorithms are neither racist nor sexist. https://itif.org/publications/2020/01/27/critics-were-wrong-nist-data-shows-best-facial-recognition-algorithms, January 2020.

[49] Ninareh Mehrabi, Fred Morstatter, N. Saxena, Kristina Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. https://arxiv.org/pdf/1908.09635.pdf, 2019.

[50] Anuradha Nagaraj. Indian police use facial recognition app to reunite families with lost children. https://www.reuters.com/article/us-india-crime-children-idUSKBN2081CU, February 2020.

[51] Joshua New. How to fix the Algorithmic Accountability Act. https://datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act/, September 2019.

[52] CBC News. Toronto police admit using secretive facial recognition technology Clearview AI. https://www.cbc.ca/news/canada/toronto/toronto-police-clearview-ai-1.5462785, February 2020.

[53] Chris Nicholson. A beginner's guide to deep reinforcement learning. https://wiki.pathmind.com/deep-reinforcement-learning, December 2019.

[54] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, page 625–632, 2005.

[55] The City of Calgary Newsroom. Facial recognition to aid investigations. https://newsroom.calgary.ca/facial-recognition-to-aid-investigations/, November 2014.

[56] Government of Canada. Guide to the Canadian Charter of Rights and Freedoms. https://www.canada.ca/en/canadian-heritage/services/how-rights-protected/guide-canadian-charter-rights-freedoms.html, June 2020.

[57] Government of Canada. Collection, retention and disposal of personal information (continued). https://laws-lois.justice.gc.ca/ENG/ACTS/P-21/page-2.html#docCont, April 2021.

[58] Office of the Privacy Commissioner. Joint investigation of Clearview AI, inc. by the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta. https://www.priv.gc.ca/en/opc-acti

ons-and-decisions/investigations/investigations-into-businesses/2021/pipeda-2021-001/, February 2021.

[59] Office of the Privacy Commissioner of Canada. Summary of privacy laws in canada. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/02_05_d_15/, Jan 2018.

[60] Office of the Privacy Commissioner of Canada. Your privacy at airports and borders. https://www.priv.gc.ca/en/privacy-topics/airports-and-borders/your-privacy-at-airports-and-borders/, December 2018.

[61] Office of the Privacy Commissioner of Canada. Consent. https://www.priv.gc.ca/en/privacy-topics/collecting-personal-information/consent/, October 2019.

[62] Office of the Privacy Commissioner of Canada. PIPEDA fair information principles. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/p_principle/, May 2019.

[63] Office of the Privacy Commissioner of Canada. Clearview AI ceases offering its facial recognition technology in Canada. https://priv.gc.ca/en/opc-news/news-and-announcements/2020/nr-c_200706/, July 2020.

[64] Office of the Privacy Commissioner of Canada. OPC launches investigation into RCMP's use of facial recognition technology. https://www.priv.gc.ca/en/opc-news/news-and-announcements/2020/an_200228/, February 2020.

[65] Office of the Privacy Commissioner of Canada. Appearance before the standing committee on access to information, privacy and ethics (ETHI) on facial recognition technology. ht

tps://www.priv.gc.ca/en/opc-actions-and-decisions/advice-to-parliament/2021/parl_20210510_02/, May 2021.

[66] Taylor Owen, Derek Ruths, Stephanie Cairns, Sara Parker, Charlotte Reboul, Ellen Rowe, and Sonja Solomun. Facial recognition moratorium briefing #1: Implications of a moratorium on the use of facial recognition technology in Canada. Technical report, Center for Media, Technology, and Democracy, August 2020.

[67] Taylor Owen, Derek Ruths, Stephanie Cairns, Sara Parker, Charlotte Reboul, Ellen Rowe, and Sonja Solomun. Facial recognition moratorium briefing #2: Conditions for lifting a moratorium on public use of facial recognition technology in Canada. Technical report, Center for Media, Technology, and Democracy, August 2020.

[68] Christopher Parsons. Canada's proposed privacy law reforms are not enough: A path to improving organizational transparency and accountability. Technical report, The Citizen Lab, Munk School of Global Affairs and Public Policy, University of Toronto, April 2021.

[69] Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. In *International Conference on Learning Representations*, 2021.

[70] Lynne Peeples. What the data say about police brutality and racial bias — and which reforms might work. https://www.nature.com/articles/d41586-020-01846-z, June 2020.

[71] P. Jonathon Phillips, Fang Jiang, Abhijit Narvekar, Julianne Ayyad, and Alice J. O'Toole. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception*, 8(2), February 2011.

[72] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems 30 (NIPS)*, 2017.

[73] Tristan Péloquin. Reconnaissance faciale: la SQ pourrait acquérir une technologie controversée. https://www.lapresse.ca/actualites/justice-et-faits-divers/2020-06-22/reconnaissance-faciale-la-sq-pourrait-acquerir-une-technologie-controversee?utm_source=facebook&utm_medium=social&utm_campaign=algofb&fbclid=IwAR1DWd3MOukf5u8unOeAZFrs92HvgLrnCVFDoxcRftpZf6iEWWB6F0pTDZk, June 2020.

[74] Bosheng Qin, Letian Liang, Jingchao Wu, Qiyao Quan, Zeyu Wang, and Dongxiao Li. Automatic identification of down syndrome using facial images with deep convolutional neural network. *Diagnostics*, 10(7):166–171, February 2020.

[75] Alexander Quon. Halifax police confirm use of controversial Clearview AI facial recognition technology. https://globalnews.ca/news/6607993/halifax-police-confirm-clearview-ai-facial-recognition-technology/, February 2020.

[76] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. In *AAAI 2020 Workshop on AI Evaluation*, 2021.

[77] Ban Facial Recognition. Ban facial recognition. https://www.banfacialrecognition.com/map/, 2021.

[78] Sarah Repucci and Amy Slipowitz. Freedom in the world 2020: Democracy under siege. Technical report, Freedom House, 2021.

[79] Nani Jansen Reventlow. How Amazon's moratorium on facial recognition tech is different from IBM's and Microsoft's. https://slate.com/technology/2020/06/ibm-microsoft-amazon-facial-recognition-technology.html, June 2020.

[80] Joseph P. Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: Too bias, or not too bias? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 1, pages 1–10, 2020.

[81] Andrew Russell. RCMP used Clearview AI facial recognition tool in 15 child exploitation cases, helped rescue 2 kids. https://globalnews.ca/news/6605675/rcmp-used-clearview-ai-child-exploitation/, February 2020.

[82] Sumit Saha. A comprehensive guide to convolutional neural networks — the ELI5 way. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, December 2018.

[83] Tiago Salvador, Stephanie Cairns, Vikram Voleti, Noah Marshall, and Adam M. Oberman. Bias mitigation of face recognition models through calibration. https://arxiv.org/abs/2106.03761, 2021.

[84] Sigal Samuel. Facebook will finally ask permission before using facial recognition on you. https://www.vox.com/future-perfect/2019/9/4/20849307/facebook-facial-recognition-privacy-zuckerberg, September 2019.

[85] Teresa Scassa. The gutting of consent in Bill C-11. http://www.teresascassa.ca/index.php?option=com_k2&view=item&id=336:the-gutting-of-consent-in-bill-c-11&Itemid=80#, December 2020.

[86] Teresa Scassa. How do new data protection enforcement provisions in Canada's Bill C-11 measure up? http://www.teresascassa.ca/index.php?option=com_k

2&view=item&id=337:how-do-new-data-protection-enforcement-provisions-in-canadas-bill-c-11-measure-up?&Itemid=80, January 2021.

[87] Teresa Scassa. How might Bill C-11 affect the outcome of a Clearview AI-type complaint? https://www.teresascassa.ca/index.php?option=com_k2&view=item&id=339:how-might-bill-c-11-affect-the-outcome-of-a-clearview-ai-type-complaint?&Itemid=80, February 2021.

[88] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis and image labeling services. *Proceedings of the ACM on Human-Computer Interaction*, 3:166–171, November 2017.

[89] Abhinav Sharma. ONNX Model Zoo. https://github.com/onnx/models, 2021.

[90] Jonathan Shaw. The watchers: Assaults on privacy in America. https://www.harvardmagazine.com/2017/01/the-watchers, February 2017.

[91] Barry B Sookman, Daniel G. C. Glover, and Jade Buchanan. Exceptions from consent in PIPEDA: facial recognition, privacy and Clearview. Technical report, McCarthy Tétrault, February 2021.

[92] Henrietta Spalding, Steve Taylor, Rehana Browne, Rodney Appleyard, Phyllida Swift, and Margaret Hodge. Disfigurement in the UK. Technical report, Changing Faces, May 2017.

[93] Yuan Stevens and Ana Brandusescu. Weak privacy, weak procurement: The state of facial recognition in Canada. Technical report, Center for Media, Technology, and Democracy, McGill University, April 2021.

[94] Yuan Stevens and Sonja Solomun. Stevens and Solomun: Facial recognition technology speeds ahead as Canada's privacy law lags behind. https://ottawacitizen.com/opinion/stevens-and-solomun-facial-recognition-technology-speeds-ahead-as-canadas-privacy-law-lags-behind, March 2021.

[95] Philipp Terhörst, Jan Niklas Kolf, Florian Kirchbuchner Naser Damer, and Arjan Kuijper. Post-comparison mitigation of demographic bias in face recognition using fair score normalization. *Pattern Recognition Letters*, 140:332–338, December 2020.

[96] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.

[97] Catharine Tunney. Mall real estate company collected 5 million images of shoppers, say privacy watchdogs. https://www.cbc.ca/news/politics/cadillac-fairview-5-million-images-1.5781735, October 2020.

[98] Catharine Tunney. RCMP denied using facial recognition technology - then said it had been using it for months. https://www.cbc.ca/news/politics/clearview-ai-rcmp-facial-recognition-1.5482266, March 2020.

[99] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, page 1–7. Association for Computing Machinery, 2018.

[100] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI. *SSRN Electronic Journal*, January 2020.

[101] Mei Wang and Weihong Den. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9319–9328, 2020.

[102] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial Faces in the Wild: Reducing racial bias by information maximization adaptation network. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 692–702, 2019.

[103] Karen Weise and Natasha Singer. Amazon pauses police use of its facial recognition software. `https://www.nytimes.com/2020/06/10/technology/amazon-facial-recognition-backlash.html`, June 2020.

[104] CTV Winnipeg. From facial recognition to extra staff: High and low tech tools used to combat shoplifting in Winnipeg. `https://winnipeg.ctvnews.ca/from-facial-recognition-to-extra-staff-high-and-low-tech-tools-used-to-combat-shoplifting-in-winnipeg-1.4307648`, February 2019.

[105] Wenying Wu, Pavlos Protopapas, Zheng Yang, and Panagiotis Michalatos. Gender classification and bias mitigation in facial images. In *12th ACM Conference on Web Science*, page 106–114, July 2020.

[106] Dong Yi, Zhen Lei, Shengcai Liao, and S. Li. Learning face representation from scratch. `https://arxiv.org/abs/1411.7923`, 2014.

[107] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5697–5706, 2019.

[108] Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 609–616, 2001.

Appendix A

Additional fairness definitions

All definitions are taken from [99].

A.1 Definitions based on predicted and actual outcome

Definition A.1.1 (Predictive parity)**.** Also called “outcome test", predictive parity is satisfied if both groups have equal positive predictive values (PPVs).

$$P(Y = 1|d = 1, G = a) = P(Y = 1|d = 1, G = b) \tag{A.1}$$

where $Y = 1$ and $Y = 0$ correspond, respectively, to the events of belonging and not belonging to the positive class.

Recall that PPV corresponds to the probability that a positively predicted subject genuinely belongs to the positive class.

Definition A.1.2 (Equalized odds)**.** Known also as conditional procedure accuracy equality and disparate mistreatment, equalized odds necessitates that the two groups have equal true positive

rates (TPRs) and equal false positive rates (FPRs).

$$P(d = 1|Y = i, G = a) = P(d = 1|Y = i, G = b), \;\; i \in \{0, 1\} \tag{A.2}$$

It can be shown that if $P(Y = 1|G = a) \neq P(Y = 1|G = b)$ then a model that satisfies predictive parity cannot satisfy equalized odds.

Definition A.1.3 (Conditional use accuracy)**.** Conditional use accuracy equality is satisfied if both groups possess equal positive predictive values (PPVs) and negative predictive values (NPVs). That is, both

$$P(Y = 1|d = 1, G = a) = P(Y = 1|d = 1, G = b) \tag{A.3}$$

and

$$P(Y = 0|d = 0, G = a) = P(Y = 1|d = 0, G = b) \tag{A.4}$$

must be satisfied.

Definition A.1.4 (Overall accuracy equality)**.** Overall accuracy equality requires equal prediction accuracies for both groups.

$$P(d = Y, G = a) = P(d = Y, G = b) \tag{A.5}$$

In other words, the probability that a subject from either the positive or negative class is assigned to its correct class must be the same for both groups.

Definition A.1.5 (Treatment equality)**.** Treatment equality is satisfied if the two groups have equal ratios of false positives and false negatives.

$$\frac{FN}{FP}(a) = \frac{FN}{FP}(b) \tag{A.6}$$

A.2 Definitions based on predicted probabilities and actual outcome

Definition A.2.1 (Balance for positive class)**.** Balance for the positive class is satisfied if the average predicted score s for subjects truly in the positive class is equal for both groups.

$$E(S|Y = 1, G = a) = E(S|Y = 1, G = b) \tag{A.7}$$

Definition A.2.2 (Balance for negative class)**.** Similarly, balance for the negative class is satisfied if the average predicted score s for subjects truly in the negative class is equal for both groups.

$$E(S|Y = 0, G = a) = E(S|Y = 0, G = b) \tag{A.8}$$

A.3 Similarity-based measures

Definition A.3.1 (Causal discrimination)**.** Here, fairness is achieved if the model produces the same classification for any two subjects with the exact same non-sensitive features X. In other words, two subjects who differ only by which group G they belong to should be assigned to the same class.

$$(X_m = X_f \wedge G_m \neq G_f) \rightarrow d_m = d_f \tag{A.9}$$

Definition A.3.2 (Fairness through unawareness)**.** Fairness through unawareness is achieved if no sensitive attribute is used in the decision-making process.

In theory, the classification outcomes should be the same for any subjects i & j who have the same non-sensitive features X:

$$X : X_i = X_j \rightarrow d_i = d_j \tag{A.10}$$

However, in practice, other features often unintentionally serve as a proxy for the sensitive attribute (e.g. neighbourhood serving as a proxy for race), thereby affecting the classification outcomes of subjects from different demographic groups.

Definition A.3.3 (Fairness through awareness)**.** Fairness through awareness requires that similar individuals be classified similarly. Here, similarity is defined by D, a distance metric between the distribution of model outputs. Given a set of subjects V, we define $K : V \times V \rightarrow R$ as being the distance between subjects. To satisfy fairness through awareness, D must be no more than K. That is, fairness is achieved if for any subjects x and y:

$$D(M(x), M(y)) \leq K(x, y) \tag{A.11}$$

where $M : V \rightarrow \delta A$ is a mapping from the set of subjects to the probability distribution of the model outcomes.

Appendix B

Bias mitigation methods - additional details

B.1 Adversarial training: AGENDA [22]

Recall that Dhar et al.'s [22] AGENDA contains four components: a pre-trained network P that generates a face embedding f_{in}; a generator model M, which takes f_{in} as input and outputs a lower dimensional f_{out}; an identity classifier C; and a set of K gender predictors $E_1, E_2...E_K$.

Training for AGENDA occurs over four key stages. **Stage 1** involves initializing and training the generator model M and the identity classifier C. The face embedding f_{in} (outputted from the pre-trained network) is passed through M to generate f_{out}. M and C are then trained using

$$L_{class}(\phi M, \phi C) = -y_{id} \log (\hat{y}_{id}) \tag{B.1}$$

where $\hat{y}_{id} = C(f_{out}, \phi C)$, and where ϕM and ϕC are the weight of M and C.

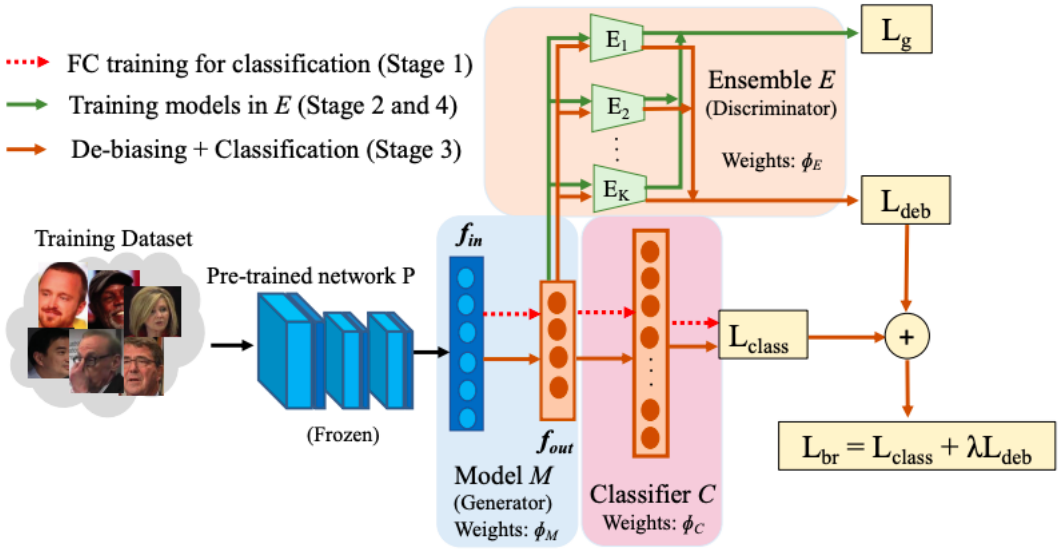

Figure B.1: AGENDA architecture. Figure source: [22]

Stage 2 initializes and trains the gender predictor model E. The outputs f_{out} generated by M are passed into E, which is then trained to classify gender using

$$L_g(\phi M, \phi E) = \sum_{k=1}^{K} L_g^{E_k} \tag{B.2}$$

where

$$L_g^{E_k}(\phi M, \phi E_k) = -y_g \log y_g^{(k)} - (1 - y_g) \log (1 - y_g^{(k)}) \tag{B.3}$$

Here, y_g is the binary gender label for the input face embedding f_{out} and $y_g^{(k)}$ is the softmax output of E_k.

Next, M and C are updated during **Stage 3**. M is trained to generate face embeddings f_{out} that can be used to predict identity but not gender. First, f_{out} is fed into C and the classifier training loss L_{class} is computed. Next, f_{out} is fed into each model in E and used to calculate gender probability scores $o_k^{male}, o_k^{female} = E_k(f_{out}, \phi E_k)$. If E is gender-agnostic, as desired, o_k^{male} and o_k^{female} would both equal 0.5.

The adversarial loss for the model E_k can then be defined as

$$L_a^{E_k}(\phi M, \phi E_k) = -(0.5 * \log(o_k^{male}) + 0.5 * \log(o_k^{female})) \tag{B.4}$$

Since the aim is to make f_{out} gender-agnostic with respect to multiple gender prediction models, they compute the adversarial loss for each model E_k and select the model with the maximum loss:

$$L_{deb}(\phi M, \phi E) = \max\{L_a^{E_k}(\phi M, \phi E_k)|_{k=1}^{K}\} \tag{B.5}$$

They then combine L_{class} and L_{deb} into

$$L_{br}(\phi C, \phi M, \phi E) = L_{class}(\phi C, \phi M) + \lambda L_{deb}(\phi M, \phi E) \tag{B.6}$$

which is used to train M and C, while ϕE remains unchanged. During training, ϕC is updated via the gradients of L_{class}, while ϕM is updated via the gradients of both L_{class} and L_{deb}.
Finally, in **Stage 4**, E is re-trained to predict gender based on f_{out}. Stage 3 (training M to fool E and training C) and stage 4 (training E) are alternated until training is complete.

B.2 Reinforcement learning: Race balanced network

Wang and Deng's [101] race balanced network (RL-RBN) uses a Markov decision process to identify optimal margins for each demographic group. This process consists of the following steps: Deep Q-learning—an important component of reinforcement learning—is used to train an agent to create an adaptive margin policy for subjects of colour. Deep Q-learning trains the agent to learn which actions to take to maximize a reward function—in this case, the reward is related to the distances between each race's face embeddings for intra- and inter-subject pairs. In particular, the reward is based on the difference, $B_{inter}^{g,w}$, between the inter-subject distances for white subjects

and for subjects from each other race $g \in \{$ Indian, Asian, African $\}$.

$$B_{inter}^{g,w} = |d_{inter}^{g} - d_{inter}^{w}| \tag{B.7}$$

where

$$d_{inter}^{g} = \frac{1}{N_g} \sum_{i=1}^{N_g} \max_{k=1:N_g, k \neq i} cost(c_k, c_i) \tag{B.8}$$

where N_g is the number of subjects in group g, and c_i is computed by taking the mean face embedding of the i-th subject in group g.

The reward is likewise based on the intra-subject equivalent to $B_{inter}^{g,w}$:

$$B_{inter}^{g,w} = |d_{intra}^{g} - d_{intra}^{w}| \tag{B.9}$$

where

$$d_{intra}^{g} = \frac{1}{N_g} \sum_{i=1}^{N_g} \frac{1}{|I_i|} \sum_{x_j \in I_i} cos(x_j, c_i) \tag{B.10}$$

where I_i is the set of all images associated with subject i.

Once the optimal adaptive margin policy is determined, it is used to train the FR model in such a way that each race has similar intra- and inter-subject distances.

B.3 Domain adaptation

Wang et al. [102] proposed IMAN, a deep information maximization network inspired by unsupervised domain adaption (UDA). UDA involves applying a network trained on a labeled source domain to an unlabeled target domain. Here, the labeled source domain is a set of facial images of white subjects whose identities (i.e. labels) are known, and the unlabeled target domain is a set of images of non-white subjects whose identities are unknown. Standard UDA methods use the source domain's classifier to estimate labels for the target domain and are thus incompatible with facial recognition, since the source and target domains do not share labels (i.e. identities). The

authors proposed an alternative method to generate pseudo-labels for the target domain: a clustering algorithm that retains only strongly-connected groups of images with similar cosine similarity scores (as computed by the network). Pseudo-labels were thereby bestowed upon these clusters and their associated images.

The authors noted, however, that due to the inadequacy of these pseudo-labels, further steps are required to ensure that the network performs well on the target domain. To that end, they introduced a mutual information loss, which uses all target images (not only those retained clusters) to learn larger decision margins for—and thereby improve the network's classification abilities on—the target domain.

The team tested three versions of their approach: transferring information about white faces to African, Asian, and Indian target domains. As described in the main text, IMAN performed better (in terms of true positive rates and ROC curves) than other models when applied to Indian, Asian, and African subjects.

B.4 Fair Template Comparison

Terhörst et al. [96] tested two different losses, one based on group fairness (i.e. fairness between demographic groups) and the other based on individual fairness (i.e. the similar treatment of similar individuals) for their post-hoc neural network. They claimed that the second loss, which we present below and which we employed for our tests (see Chapter 4), resulted in a more substantial reduction in bias (up to 52.67% vs up to 41.22%). For an image pair with a true class of y (i.e. a genuine pair or an imposter pair containing images of two different individuals), and a predicted class of $\hat{y}$, they computed the loss as follows:

$$L = (1 - \lambda)H(y) + \lambda f(y, \hat{y}) + \gamma l_2 \tag{B.11}$$

where H is a binary cross-entropy function, and f is a penalization term based either on group or individual fairness. The fairness parameter λ controls the trade-off between higher accuracy and a higher degree of fairness. Finally γl_2 is an l_2 regularization to limit overfitting.

Given a set of demographic groups G, they let S_g such that $g \in G$ be a set of face embeddings that belong to group g. Then, for the case of individual fairness, they set

$$f = \sum_{i,j \in G} \frac{1}{|S_i| \cdot |S_j|} \sum_{\substack{(y_k, \hat{y}_k) \in S_i \\ (y_l, \hat{y}_l) \in S_j}} \delta(y_k, y_l)(\hat{y}_k - \hat{y}_l)^2 \tag{B.12}$$

where

$$\delta(y_k, y_l) = \begin{cases} 1, & \text{if } x_k = x_l \\ 0, & \text{if } x_k \neq x_l \end{cases} \tag{B.13}$$